NASA TECHNICAL
MEMORANDUM

NASA TM X-53024

March 17, 1964

324 p

N64-20949 → N64-20966

Code 3 - Cat 01

NASA TM X-53024

PROGRESS REPORT NO. 5

STUDIES IN THE FIELDS OF SPACE FLIGHT AND GUIDANCE THEORY

Sponsored by Aero-Astrodynamics Laboratory

NASA

George C. Marshall
Space Flight Center,
Huntsville, Alabama

NASA-GEORGE C. MARSHALL SPACE FLIGHT CENTER

TECHNICAL MEMORANDUM X-53024

March 17, 1964

PROGRESS REPORT NO. 5
on Studies in the Fields of
SPACE FLIGHT AND GUIDANCE THEORY
Sponsored by Aero-Astrodynamics Laboratory of
Marshall Space Flight Center

ASTRODYNAMICS AND GUIDANCE THEORY DIVISION
AERO-ASTRODYNAMICS LABORATORY

NASA-GEORGE C. MARSHALL SPACE FLIGHT CENTER

TECHNICAL MEMORANDUM X-53024

PROGRESS REPORT NO. 5
on Studies in the Fields of
SPACE FLIGHT AND GUIDANCE THEORY

Sponsored by Aero-Astrodynamics Laboratory of
Marshall Space Flight Center

ABSTRACT

This paper contains progress reports of NASA-sponsored studies in the areas of space flight and guidance theory. The studies are carried on by several universities and industrial companies. This progress report covers the period from July 18, 1963, to December 18, 1963. The technical supervisor of the contracts is W. E. Miner, Deputy Chief of the Astrodynamics and Guidance Theory Division, Aero-Astrodynamics Laboratory, George C. Marshall Space Flight Center.

TABLE OF CONTENTS

TABLE OF CONTENTS (CONT'D)

NASA-GEORGE C. MARSHALL SPACE FLIGHT CENTER

TECHNICAL MEMORANDUM X-53024

~~PROGRESS REPORT NO. 5~~
on Studies in the Fields of
SPACE FLIGHT AND GUIDANCE THEORY

Sponsored by Aero-Astrodynamics Laboratory of
Marshall Space Flight Center

SUMMARY

This paper contains progress reports of NASA-sponsored studies in the areas of space flight and guidance theory. The studies are carried on by several universities and industrial companies. This progress report covers the period from July 18, 1963 to December 18, 1963. The technical supervisor of the contracts is W. E. Miner, Deputy Chief of the Astrodynamics and Guidance Theory Division, Aero-Astrodynamics Laboratory, George C. Marshall Space Flight Center.

INTRODUCTION

This report contains fifteen papers, the subject matter of which lies in the areas of guidance and space flight theory. These papers were written by investigators employed at agencies under contract to Marshall Space Flight Center.

This report is the fifth of the "Progress Reports" and covers the period from July 18, 1963 to December 18, 1963. Information given in Progress Reports 1 through 4 will not be repeated here.

The agencies contributing and their fields of major interest are:

Field of Interest	Agency
Calculus of Variations	Grumman Aircraft Engineering Corp. Auburn University Analytical Mechanics Associates General Electric Company
Impulse Orbit Transfer	North American Aviation, Inc.
Matrix Operations	University of Kentucky
Large Computer Exploitation	University of North Carolina Northeast Louisiana State College Georgia Institute of Technology
Stability	Martin Merrita Company Minneapolis Honeywell Regulator Co.
Low Thrust Trajectories	Aeronutronics (Ford)

The objectives of this introduction are to review and summarize the contributions of each agency.

The first paper by McGill and Kenneth of Grumman Aircraft Engineering Corporation describes a computational procedure for obtaining the solution to a nonlinear two-point boundary value differential equations problem. The procedure is based on a generalization of the Newton-Raphson technique as a contraction mapping in a suitably defined metric space. Thus, the solution is arrived at through a sequence of solutions to related systems of linear differential equations rather than the usual sequence of approximate solutions to the nonlinear problem. Analytical estimates to convergence properties are not given, but the numerical results given indicate that convergence might be rapid in many cases if some feel for the character of the solution furnishes a reasonable arbitrary initial function. It appears that the procedure may develop into an economical tool for isolating extremal trajectories.

The second paper by Harmon and Shaw of Auburn University develops a system of differential equations that defines optimum reentry trajectories corresponding to a specified hardware arrangement. The attitude of the reentry vehicle is assumed to be controlled in one degree of freedom (yaw angle of attack) by means of an offset center of gravity and roll jets. The remaining degrees of freedom in attitude are constrained by relationships resulting from assumed steady state solutions. The minimized variable is the time integral of the total drag squared. The equations are developed in some detail and presented with instructions in a form amenable to coding on a high speed computer, so that numerical studies could be carried out to determine how much the minimized variable is increased and target acquisition ability is decreased by these hardware constraints.

Some work toward the analytical or direct derivation of guidance functions is contained in the third paper by Kelly of Analytical Mechanics Associates, Inc. The approach is an application of perturbation theory to the Euler-Lagrange equations with expansions truncated after the second order terms. The theory is applied to a simple problem for illustrating the theory and providing some information on the contribution of second order terms compared to that of first order terms. Some difficulties may be encountered when the procedures as given are applied to our actual problems. The on-board storage of the nominal optimum trajectory is one, and the required closed form solution to the differential equations is another. These difficulties can probably be overcome. An approach which is theoretically very similar to Dr. Kelly's has been carried out by R. Silber of Southern Illinois University. The method is being coded and evaluated in-house for optimum flight assuming a spherical earth and produces a guidance function approximation in the usual polynomial form.

The fourth paper by Pines of Analytical Mechanics Associates, Inc. suggests a possible basis for iterative solutions to the two-point boundary value problem associated with trajectory optimization by indirect methods. He proposes to use the impulsive thrust solution with the constants of the motion that he derives, and determine initial values of the adjoint variables as limits of the values for the finite thrust case as thrust increases without bound. This would furnish the first guess in some iterative process for solving a finite thrust case. The desired finite thrust solution might be arrived at through some sequence of such iterative solutions. However, for some missions it may be nearly as difficult to solve the optimization problem with impulsive thrust as it is to solve the original finite thrust problem. Numerical evaluations of the method are yet to be made.

Mr. Cavoti of General Electric in the fifth paper treats a simplified problem of optimum retrothrust in an inverse square gravitational field. The principal restrictions are a thrust direction always tangent to the flight path, and end-conditions independent of range and time. Under these conditions, he finds that the optimum thrust magnitude program for bounded variable retrothrust might consist of subarcs of minimum, variable intermediate, or maximum thrust. A closed form solution is found for the intermediate thrust case that implies a constant velocity magnitude over such an arc. There is some question as to whether these results can be helpful toward the solution of the unrestricted problem.

The sixth paper by Gentry Lee of North American Aviation, Inc. shows - for the rather restricted subfamily of transfer orbits characterized as coplanar-elliptical and of equal angular momentum - that there exists a specific family of two impulse transfers that use no more impulse than a one impulse transfer at the intersection of the two orbits. The results of this study represent a step toward the solution of the n-impulse transfer problem in which it has been conjectured that an n-impulse transfer would require less fuel than any transfer using fewer impulses.

The seventh paper is written by D. F. Bender of North American Aviation, Inc. In this paper as in the companion paper by Gentry Lee, a comparison is made between one impulse and two impulse transfers between orbits of a rather restricted family. This family consists of nearly tangent coplanar elliptical orbits. It is found that over a narrow range of orbit shapes for these shallowly intersecting orbits, one impulse and optimum two impulse transfers require practically identical total impulses.

The eighth paper is written by the University of Kentucky Team. It presents a matrix method for representing the general cubic

$$\sum_{i,j,k} \alpha_{ijk} x_i x_j x_k$$

and for finding the coefficients of this cubic subjected to the transformation $x_i=y_i+\beta_i$, $i=1,2,\ldots,n$. This procedure enables one to compute the coefficients of the new cubic, in $y_i y_j y_k$, in any order and to apply approximation techniques to the result.

The ninth paper, written by Shigemichi Suzuki of the University of North Carolina, describes new iterative algorithms as alternatives to solving problem (a) as set forth in Progress Report No. 4. Problem (a) is: given a fixed form for the ratio of linear combinations of known functions, coefficients are sought such that the maximum deviation over a finite point set is a minimum (Tchebycheff). The auxiliary functions are optimized under the constraints given previously. These algorithms may provide more effective means of solving the problem than those presented in Progress Report No. 4. These methods will also be evaluated by MSFC on the problem of generating steering and time of cutoff functions.

The tenth paper, "Inverse Estimation" by G. W. Adkins, also of the University of North Carolina, presents a novel approach for empirically fitting guidance functions. The procedure utilizes an algorithm which specifies the control variables for a given sample of the response variables. In this process of function approximation, the role of independent and dependent variables are reversed. The procedure has not been fully evaluated, but at this time it would seem impractical for use.

The eleventh paper written by the group from Northeast Louisiana State College describes a technique for obtaining the numerical values of a function which yields an error in the sense of least squares that is equal to a specified tolerance. In the notation of the paper the least squares error is defined as

$$E = \|\overline{x} - \sum_{i=0}^{N} A_i \; \overline{\varphi_i}\|^2$$

where the vectors $\overline{x_i}$ and $\overline{\varphi_i}$ have n components corresponding to the number of points used. Starting with the expression for E, a method is developed for determining the n numerical values of the components of the vector φ_{N+1} such that

$$\|\overline{x} - \sum_{i=0}^{N+1} A_i \; \varphi_i\|^2 = \delta \text{ (a specified tolerance).}$$

More effort is needed to determine a suitable functional form for the vector $\overline{\varphi_{N+1}}$ or to find a use for the numerical values of its components.

The twelfth paper, prepared by the group at Georgia Institute of Technology, describes a method of obtaining least squares estimates of multivariable polynomials. By using a particular polynomial form, called a balanced polynomial, the "step procedure" method yields least squares estimates while reducing the order of the matrix to be inverted. The method has not been applied but shows promise.

The thirteenth paper was written by D. L. Lukes of Minneapolis-Honeywell Regulator Company. In this paper it is assumed that some open loop (reference) trajectory and the required control have been determined for a given dynamical system. The problem of extending the control to a neighborhood of the reference trajectory to obtain a feedback control that will drive the system to the desired final state is investigated. The technique used is based on the construction of a Lyapunov function defined in some neighborhood of the reference trajectory. This technique differs from the classical linearization of the system equations. Furthermore, stability is assured. It appears that the same technique may be applied to an n-dimensional system with a vector control function.

The fourteenth paper by H. Hermes of Martin Marietta Company discusses "Controllability for Linear and Nonlinear Systems." The idea of complete controllability for linear control systems was first introduced and exploited by R. E. Kalman, Y.C. Ho, and K. S. Narendra. In this paper Dr. Hermes extends this concept to nonlinear systems with the control appearing linearly. The first part of the paper summarizes the work of the above authors. The second part is concerned with the extension of the concept to systems of the form

$$\dot{\chi}(t) = g(t, (\chi, t)) + H(t, \chi(t))\ u(t)$$

where g is an n-vector, H is an $n \times r$ matrix, and u is a finite valued measurable control vector. An argument is presented with regard to what should be meant by complete controllability of the system stated above.

On the basis of the characteristics believed desirable for this concept, a criterion is stated and it is shown to satisfy the selected characteristics. How to extend the concept to other nonlinear systems remains an open question.

The last paper of this progress report concerns low thrust trajectories. It is written by D. P. Johnson and L. W. Stumpf of Aeronutronic Division of Ford Motor Company. The paper presents a complete second-order solution for the case in which the thrust vector makes an arbitrary but constant thrust angle with the radius vector. The solution presented is constrained to leave a circular orbit. Further work should be done to relax this condition and also on selecting optimum thrusting angles.

Dr. Mary Payne of Republic Aviation reworked some of the material presented in her paper, "Application of the Two Fixed Center Problem to Lunar Trajectories." This work is presented under item 17 in the Table of Contents.

It will be noted again that the editors of this report do not correct any of the papers and the authors are responsible for their papers in detail.

RESEARCH DEPARTMENT

GRUMMAN AIRCRAFT ENGINEERING CORPORATION

SOLUTION OF VARIATIONAL PROBLEMS BY MEANS OF A GENERALIZED NEWTON-RAPHSON OPERATOR

By

Robert McGill
Paul Kenneth

BETHPAGE, NEW YORK

RESEARCH DEPARTMENT
GRUMMAN AIRCRAFT ENGINEERING CORPORATION
BETHPAGE, NEW YORK

SOLUTION OF VARIATIONAL PROBLEMS BY MEANS OF A GENERALIZED NEWTON-RAPHSON OPERATOR

by

Robert McGill
Paul Kenneth

A

Summary

20951

This paper presents the development of an indirect method for solving variational problems by means of an algorithm for obtaining the solution to the associated nonlinear two-point boundary value problem. The method departs from the usual indirect procedure of successively integrating the nonlinear equations and adjusting arbitrary initial conditions until the remaining boundary conditions are satisfied. Instead, an operator is introduced which produces a sequence of sets of functions which satisfy the boundary conditions but in general do not satisfy the nonlinear system formed by the state equations and the Euler-Lagrange equations. Under appropriate conditions this sequence converges uniformly and rapidly (quadratically) to the solution of the nonlinear boundary value problem.

The computational effectiveness of the algorithm is demonstrated by three numerical examples. Author

INTRODUCTION

The mathematical theory used for the study of optimization problems is the Calculus of Variations. Application of this theory to meaningful models of physical situations generally results in a mathematical representation of the solution which requires some numerical technique to effect solutions of use to the engineer. Since the major computational device available today is the high speed digital computer, e.g., the IBM 7094, an a priori requirement for a numerical algorithm is that it be systematically adaptable to high speed digital computation. For the Calculus of Variations there are two general numerical approaches; the Direct Methods, and the Indirect Methods. The direct methods proceed by solving a sequence of nonoptimal problems with the property that each successive set of solution functions yields an improved value for the functional being optimized. An example of such a procedure is the Method of Gradients which has been applied to a variety of problems with considerable success. The indirect methods are concerned to find by numerical means a set of functions which satisfy the necessary conditions for an extremal, i.e., the Euler-Lagrange differential equations. These necessary conditions and boundary conditions form a <u>nonlinear</u> boundary value problem and it is here that the numerical difficulty arises. The usual approach to this problem is the systematic variation of arbitrarily chosen initial conditions until the remaining boundary conditions are met. This technique has proved largely unsuccessful owing to increased dimensionality of the interesting problems and to the sensitivity of boundary conditions to small changes in initial conditions. In lieu of this an algorithm has been developed which proceeds by solving a sequence of <u>linear</u> boundary value problems such that the sequence of solutions converges to the solution of the nonlinear problem. Since the linear boundary value problem is easily handled numerically the algorithm is readily adaptable to high speed digital computation.

In what follows we shall discuss this approach in some detail including a discussion of the numerical application. This is followed by three numerical examples to illustrate the computational effectiveness of the method.

THE GENERALIZED NEWTON-RAPHSON OPERATOR

We are concerned with nonlinear operator equations of the following form

$$BX = 0$$

where X is an element of an appropriate metric space S and B is a nonlinear operator which maps S into itself.

For the case of the nonlinear two-point boundary value problems of interest herein the operator equation $BX = 0$ is given by the following system of nonlinear differential equations and boundary conditions

$$\dot{X} - F(X,t) = 0 \quad , \quad t \in [t_0,t_f]$$

$$\begin{array}{ll} x^{(1)}(t_0) = x_0^{(1)} & x^{(1)}(t_f) = x_f^{(1)} \\ \vdots \quad \vdots & \vdots \quad \vdots \\ x^{(\frac{N}{2})}(t_0) = x_0^{(\frac{N}{2})} & x^{(\frac{N}{2})}(t_f) = x_f^{(\frac{N}{2})} \quad , \end{array}$$

where

$$X = \left(x^{(1)},\ldots,x^{(N)}\right)$$

$$F = \left(f^{(1)},\ldots,f^{(N)}\right)$$

$$f^{(i)} = f^{(i)}\left(x^{(1)},\ldots,x^{(N)},t\right) , \qquad i = 1,\ldots,N .$$

The metric space S is given by

$$S = \left\{ X(t): \ x^{(i)}(t) \ \text{ is continuous on } \ [t_0, t_f] \ , \ i = 1, \ldots, N \right\} ,$$

with the metric

$$\rho(X_1, X_2) = \sum_{i=1}^{N} \max_{t} |x_2^{(i)}(t) - x_1^{(i)}(t)| \quad , \quad X_1, X_2 \in S \ .$$

We define an operator A on S by $X_{n+1} = AX_n$, $n = 0,1,\ldots$; X_0 arbitrary in S,

$$\dot{X}_{n+1} = J(X_n, t)[X_{n+1} - X_n] + F(X_n, t)$$

$$\begin{array}{llll} x_n^{(1)}(t_0) = x_0^{(1)} & & x_n^{(1)}(t_f) = x_f^{(1)} \\ \quad \vdots \qquad \vdots & & \quad \vdots \qquad \vdots \\ x_n^{(\frac{N}{2})}(t_0) = x_0^{(\frac{N}{2})} & & x_n^{(\frac{N}{2})}(t_f) = x_f^{(\frac{N}{2})} \end{array}$$

$$n = 0,1,2,\ldots ,$$

where $J(X,t)$ is the Jacobian matrix of partial derivatives of the $f^{(i)}$ with respect to the $x^{(j)}$, $i = 1,\ldots,v$, $j = 1,\ldots,N$. Under appropriate conditions the sequence $\{X_n\}$ converges strongly to the solution X^* of the operator equation $BX = 0$, i.e., $\lim_{n \to \infty} \rho(X_n, X^*) = 0$, where X^* is the solution of the nonlinear boundary value problem. The metric ρ implies uniform convergence for each of the component functions $x^{(i)}(t)$ of $X(t)$.

The operator A is called the Generalized Newton-Raphson operator since it may be obtained from a direct generalization of the Newton-Raphson sequence for finding roots of scalar equations. For the scalar case the operator equation $BX = 0$ becomes

$$f(x) = 0$$

and the sequence defining A becomes

$$0 = f'(x_n)[x_{n+1} - x_n] + f(x_n) \quad , \quad n = 0,1,2,\ldots .$$

The appropriate metric-space S is the scalar field with the usual metric. As before, $x_{n+1} = Ax_n$, $n = 0,1,2,\ldots,$ and x_0 is an approximate solution of $f(x) = 0$. As can be seen from the scalar application the basic concept involved is geometric; a curve is sequentially replaced by its tangent line, i.e., the nonlinear problem is replaced by a sequence of linear problems. Since there is a well developed structure for linear problems, e.g., superposition for systems of linear differential equations, the algorithm becomes computationally attractive. In addition, since the linear two-point boundary value problem can be reduced to repeated numerical integration of initial value problems, the method is readily adapted to high speed automatic machine computation.

This algorithm was apparently first suggested for boundary value problems by Hestenes (Ref. 1) who called it "Differential Variations," and later further developed by Bellman and Kalaba (Ref. 2) who refer to the technique as "Quasilinearization."

Kalaba gives a convergence proof (Ref. 2), based on monotonicity and convexity arguments, for the case of a single second order differential equation with two-point boundary conditions. A convergence proof for N dimensional systems was given by McGill and Kenneth (Ref. 3). The latter proof proceeds by establishing sufficient conditions for the operator A to be a contraction of a complete metric space into itself. The desired results then follow from the Contraction Mapping Principle (Ref. 4). The method is also mentioned by Kelley (Ref. 5) who remarks that computational experience with the technique is lacking.

NUMERICAL APPLICATION

In this section we present a brief description of a numerical procedure for solving the linear system. This procedure, with appropriate modifications, was used in obtaining the solutions to the numerical examples included in this report.

At the $n+1^{st}$ stage of the iteration we have the linear system

$$\dot{X}_{n+1} = J(X_n,t)[X_{n+1} - X_n] + F(X_n,t)$$

which is equivalent to

$$\dot{X} = C(t)X(t) + D(t) \quad , \quad t \in [t_0,t_f]$$

$$X = (x_1,\ldots,x_N)$$

$$\begin{array}{ll} x_1(t_0) = x_{10} & x_1(t_f) = x_{1f} \\ \vdots \quad\quad \vdots & \vdots \quad\quad \vdots \\ x_{\frac{N}{2}}(t_0) = x_{\frac{N}{2}0} & x_{\frac{N}{2}f}(t_f) = x_{\frac{N}{2}f} \end{array}$$

Generate by numerical integration a set $\left\{X^{(\frac{N}{2}+i)}(t)\right\}$, $i = 1,\ldots,\frac{N}{2}$, of solutions of the homogeneous system $\dot{X} = C(t)X(t)$ with initial conditions

$$X^{(\frac{N}{2}+1)}(t_0) = (0,0,\ldots,0,x_{\frac{N}{2}+1} = 1,0,\ldots,0)$$

$$X^{(\frac{N}{2}+2)}(t_0) = (0,0,\ldots,0,x_{\frac{N}{2}+2} = 1,0,\ldots,0)$$

$$\vdots$$

$$X^{(N)}(t_0) = (0,0,\ldots,0,\ldots,0,1) \ .$$

Generate a particular solution $X^{(P)}(t)$ of the nonhomogeneous system $\dot{X} = C(t)X(t) + D(t)$ with initial conditions

$$X^{(P)}(t_0) = (x_{10},x_{20},\ldots,x_{\frac{N}{2}0},K_1,K_2,\ldots,K_{\frac{N}{2}}) \ ,$$

where K_i, $i = 1,\ldots,\frac{N}{2}$, are arbitrary, e.g., $K_1 = K_2 = \ldots = K_{\frac{N}{2}} = 0$. They should, however, following a suggestion by Richard Bellman, be chosen to preserve numerical precision in solving the $\frac{N}{2}$ simultaneous linear equations given below. The solution $X(t)$ of the nonhomogeneous system with the prescribed boundary conditions is then given by

$$X(t) = c_{\frac{N}{2}+1}X^{(\frac{N}{2}+1)}(t) + c_{\frac{N}{2}+2}X^{(\frac{N}{2}+2)}(t) + \ldots + c_N X^{(N)}(t) + X^{(P)}(t) \ ,$$

where the $\frac{N}{2}$ constants $c_{\frac{N}{2}+i}$, $i = 1, \ldots, \frac{N}{2}$, are determined from the boundary conditions at $t = t_f$ by the solution of $\frac{N}{2}$ simultaneous linear equations.

For the purpose of conserving rapid access storage and also as a check on the solution of the linear system the solution $X(t)$ was not obtained from the linear combination given above. Rather it was calculated by once more integrating the nonhomogeneous system $\dot{X} = C(t)X(t) + D(t)$ with initial conditions

$$X(t_0) = (x_{10}, x_{20}, \ldots, x_{\frac{N}{2}0}, c_{\frac{N}{2}+1} + K_1, c_{\frac{N}{2}+2} + K_2, \ldots, c_N + K_{\frac{N}{2}}) \ .$$

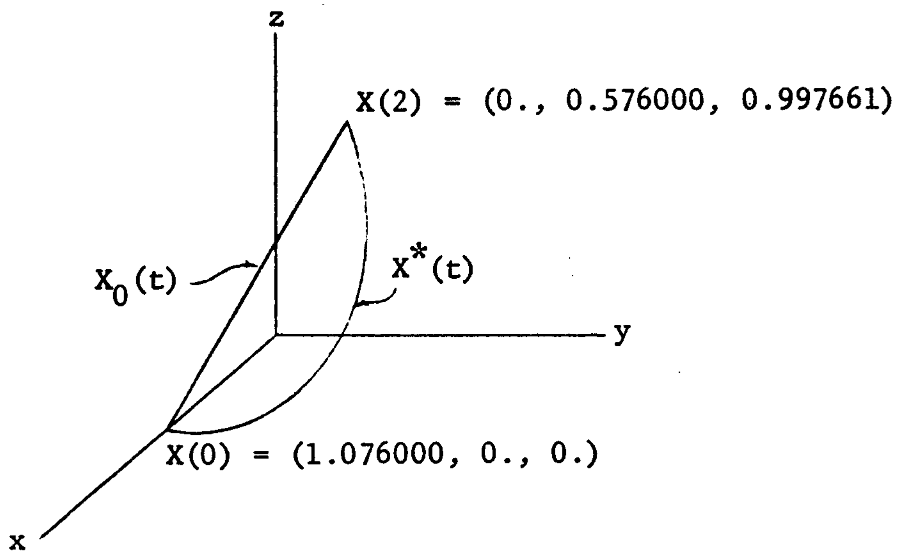
The latter procedure requires the storage of only the final values of the vectors $\left\{X^{(\frac{N}{2}+i)}\right\}$, $i = 1, \ldots, \frac{N}{2}$, and the final value of $X^{(P)}$, the particular solution.

ORBITAL INTERCEPT EXAMPLE

The first example although not an optimization problem serves to illustrate the application of the algorithm to a given nonlinear boundary value problem.

The problem solved is that of determining the free fall path which a space vehicle must follow in transferring from a specified position three hundred miles above the earth to another specified position six hundred miles above the earth, with a fixed transit time. The vehicle is assumed to be in coasting flight and the perturbing effect of the moon is included. A schematic diagram of the problem is shown below where $X_0(t) = \left(x_0(t), y_0(t), z_0(t)\right)$, the starting vector, is of the simplest possible form, namely, the straight line joining the two points in space, $X^*(t)$ is the solution vector.

The unit of length is taken to be the radius of the earth and the principal gravitational constant is normalized to one. This results in a time unit of 805.46 seconds.



Schematic Diagram

The sixth order nonlinear system and two point boundary conditions which furnish the mathematical description of the problem are given by

$$\ddot{x} = -K\frac{x}{r^3} + K_M\left(\frac{x_M - x}{\delta^3} - \frac{x_M}{r_M^3}\right)$$

$$\ddot{y} = -K\frac{y}{r^3} + K_M\left(\frac{y_M - y}{\delta^3} - \frac{y_M}{r_M^3}\right) \quad ; \quad t \in [0,2]$$

$$\ddot{z} = -K\frac{z}{r^3} + K_M\left(\frac{z_M - z}{\delta^3} - \frac{z_M}{r_M^3}\right)$$

$$x(0) = 1.076000 \qquad x(2) = 0.$$

$$y(0) = 0. \qquad y(2) = 0.576000$$

$$z(0) = 0. \qquad z(2) = 0.997661$$

$$r = [x^2 + y^2 + z^2]^{\frac{1}{2}}$$

$$r_M = [x_M^2 + y_M^2 + z_M^2]^{\frac{1}{2}}$$

$$\delta = \left[(x_M - x)^2 + (y_M - y)^2 + (z_M - z)^2\right]^{\frac{1}{2}}$$

For simplicity the lunar coordinates, x_M, y_M, z_M, are assumed constant.

The time interval $[0,2]$ was divided into 100 parts and the necessary numerical integrations carried out by means of a high speed digital computer (IBM 7094) to an accuracy of seven significant figures. The results are exhibited in Table 1 where for brevity only six points in time are shown. $X_0(t)$ is the linear starting function; $X_1(t)$ is the first mapping; $X_2(t)$ is the second mapping, etc.; and $X^*(t)$ results from the integration of the actual nonlinear equations with the initial velocities,

$$\dot{x}(0) = 0.101637$$

$$\dot{y}(0) = 0.472285$$

$$\dot{z}(0) = 0.818022 \ ,$$

obtained from the final iterate.

The sequence $\{X_n\}$ converged, within the accuracy of our computations, in three iterations with:

$$\rho(X_1,X_0) = 0.480116$$

$$\rho(X_2,X_1) = 0.133753$$

$$\rho(X_3,X_2) = 0.004375$$

$$\rho(X_4,X_3) = 0.000004 \ ,$$

where

$$\rho(X_{n+1},X_n) = \max_t|x_{n+1}(t) - x_n(t)| + \max_t|y_{n+1}(t) - y_n(t)|$$

$$+ \max_t|z_{n+1}(t) - z_n(t)| \ .$$

As a further check on the over-all accuracy the perturbing force was set to zero and the final value of the magnitude of the initial velocity was compared with that obtained by the closed form solution for the two-body problem. Within the accuracy of our computations these values were identical.

We note that we have simply and rapidly produced the numerical solution to a simple orbit determination problem, viz., given the position of a body at two distinct times, determine the time varying orbital elements of the body in the presence of perturbing forces. Solutions have also been produced even when the two points are exactly 180 degrees apart. In this case the straight line could not be used as a starting function since it is singular. However, a simple triangular path was sufficient to produce the characteristic rapid convergence.

LUNAR DESCENT EXAMPLE — MAXIMUM RANGE

A very simple variational problem was chosen for the second numerical example. This problem concerns the maximization of the translational range of a lunar vehicle during descent to rest from a hovering condition 1000 ft above the lunar surface. The time for the maneuver was fixed at 2.062 minutes.

For the purpose of generating this numerical example the following simplifying assumptions were made:

Constant thrust acceleration
Uniform gravitational field
Analysis restricted to two dimensions.

The problem then is reduced to finding the thrust steering angle time history which produces the maximum range in the given fixed time.

The mathematical description of the problem is furnished by the following Euler-Lagrange differential equations and boundary conditions

$$\dot{u} = T \frac{\lambda_u}{(\lambda_u^2 + \lambda_v^2)^{\frac{1}{2}}} = f^{(1)} \quad ; \quad t \in [t_0, t_f]$$

$$\dot{r} = T \frac{\lambda_v}{(\lambda_u^2 + \lambda_v^2)^{\frac{1}{2}}} - g_M = f^{(2)}$$

$$\dot{y} = v = f^{(3)}$$

$$\dot{\lambda}_u = - 1 = f^{(4)}$$

$$\dot{\lambda}_v = - \lambda_y = f^{(5)}$$

$$\dot{\lambda}_y = 0 = f^{(6)}$$

$$u(t_0) = u_0 \qquad u(t_f) = u_f$$

$$v(t_0) = v_0 \qquad v(t_f) = v_f$$

$$y(t_0) = y_0 \qquad y(t_f) = y_f$$

The unit of length was chosen equal to the initial altitude of 1000 ft and the local gravitational constant and vehicle mass were put equal to one. This resulted in the following normalized data for the problem:

$$\begin{array}{lll}
u_0 = 0.000 & u_f = 0.000 & x_0 = 0.000 \\
v_0 = 0.000 & v_f = 0.000 & \\
y_0 = 1.000 & y_f = 0.000 & \\
T = 5.000 & t_0 = 0.000 & \\
g_M = 1.000 & t_f = 9.000 &
\end{array}$$

This normalization resulted in a time unit of 13.70 seconds. A crude starting function $X_0(t)$ was chosen as follows:

$$\begin{aligned}
u_0(t) &\equiv 0 \\
v_0(t) &\equiv 0 \\
y_0(t) &= y_0 + \frac{y_f - y_0}{t_f - t_0}\, t \\
\lambda_{y_0}(t) &\equiv c_3 \\
\lambda_{u_0}(t) &= c_1 - t \\
\lambda_{v_0}(t) &= c_2 - c_3\, t \ ,
\end{aligned}$$

where the three constants c_1, c_2, and c_3 correspond to an arbitrary estimate that the steering angle, measured from the local horizontal, should be initially zero, equal to $\pi/2$ at $t = \frac{t_f}{2}$, and slightly less than π at $t = t_f$.

The sequence $\{X_n\}$ for this case converged uniformly to an accuracy of 5 significant figures in six iterations. The total computer time (IBM 7094) required for this problem was 18 seconds. The desired final value of the range $x_f = 100,200$ ft was obtained from

$$x_f = \int_{t_0}^{t_f} u^*(t)dt ,$$

where $u^*(t)$ results from the integration of the nonlinear state and Euler-Lagrange equations with a complete set of initial values taken from the final iterate. This final integration of the nonlinear equations also served as an over-all check on the solution.

LOW THRUST ORBITAL TRANSFER EXAMPLE — MINIMUM TIME

The third and final example concerns the problem of minimizing the transfer time of a low thrust ion rocket between the orbits of Earth and Mars. This problem involves additional complications over the previous problems, the most significant of which is the fact that the final value of the independent variable is no longer fixed.

To simplify the problem as much as possible the rocket's thrust level was assumed constant, and thus the single control variable is the thrust direction. Further, the orbits of Earth and Mars were assumed to be circular and coplanar, and the gravitational attractions of the two planets on the vehicle were neglected. The following system parameters for the low-thrust vehicle were adopted from Ref. 5:

Initial Mass, m_0	46.58 slugs
Specific Impulse	4700 sec
Propellant Consumption Rate, $\dot{m}$,	-6.937×10^{-7} slugs/sec
Thrust, T,	0.127 lb
Thrust/Initial Weight	0.9×10^{-4}

The equations of motion are given by:

Radial Velocity

$$\dot{r} = f^{(1)} = u$$

Radial Acceleration

$$\dot{u} = f^{(2)} = \frac{v^2}{r} - \frac{k}{r^2} + \frac{T \sin \theta}{m_0 + \dot{m}t}$$

Circumferential Acceleration

$$\dot{v} = f^{(3)} = - \frac{uv}{r} + \frac{T \cos \theta}{m_0 + \dot{m}t}$$

where u and v are the radial and circumferential velocities respectively; r is the radius; and θ is the thrust direction angle measured from the local horizontal. All the initial and final values of the state variables were specified, and the quantity to be minimized was t_f, the final time. Since the method as previously outlined required a fixed final time, the procedure was altered to suit the minimum time problem. What follows is a brief description of the modified procedure and a discussion of the numerical results.

The two point boundary value problem resulting from the Euler-Lagrange equations is given by

$$\dot{r} = u = f^{(1)}$$

$$\dot{u} = \frac{v^2}{r} - \frac{k}{r^2} + a(t)\,\frac{\lambda_u}{\left(\lambda_u^2 + \lambda_v^2\right)^{\frac{1}{2}}} = f^{(2)}$$

$$\dot{v} = -\frac{uv}{r} + a(t)\,\frac{\lambda_v}{\left(\lambda_u^2 + \lambda_v^2\right)^{\frac{1}{2}}} = f^{(3)}$$

$$\dot{\lambda}_r = \left(\frac{v^2}{r^2} - 2\,\frac{k}{r^3}\right)\lambda_u - \frac{uv}{r^2}\,\lambda_v = f^{(4)}$$

$$\dot{\lambda}_u = -\lambda_r + \frac{v}{r}\,\lambda_v = f^{(5)}$$

$$\dot{\lambda}_v = -2\,\frac{v}{r}\,\lambda_u + \frac{u}{r}\,\lambda_v = f^{(6)}$$

where

$$a(t) = \frac{T}{m_0 + \dot{m}t},$$

and the boundary conditions are

$t = 0$	$t = t_f$ (unspecified)
$r(0) = r_0$	$r(t_f) = r_f$
$u(0) = u_0$	$u(t_f) = u_f$
$v(0) = v_0$	$v(t_f) = v_f$

This may be written as

$$\dot{X} = F(X,t)$$

where

$$X = (x^{(1)},\ldots,x^{(6)})$$

$$F = (f^{(1)},\ldots,f^{(6)})$$

and

$$x^{(1)}(t) = r(t) \quad , \quad x^{(2)}(t) = u(t) \quad , \quad x^{(3)}(t) = v(t)$$

$$x^{(4)}(t) = \lambda_r(t) \quad , \quad x^{(5)}(t) = \lambda_u(t) \quad , \quad x^{(6)}(t) = \lambda_v(t) \; .$$

The method proceeds as before by solving the following sequence of linear two point problems

$$\dot{X}_{n+1} = J(X_n,t)[X_{n+1} - X_n] + F(X_n,t \qquad n = 0,1,\ldots ,$$

where $J(X,t)$ is the Jacobian matrix of partial derivatives of the $f^{(i)}$ with respect to the $x^{(j)}$, $i = 1,\ldots,6$, $j = 1,\ldots,6$. A starting vector, $X_0(t)$ and an estimated final time, t_{f_0}, are assumed and the sequence of linear boundary value problems is solved numerically by the procedure outlined previously, with the following boundary values:

$$\begin{array}{ll} t = 0 & t = t_{f_k} \\ x_n^{(1)}(0) = r_n(0) \ = r_0 & \\ x_n^{(2)}(0) = u_n(0) \ = u_0 & x_n^{(2)}(t_f) = u_n(t_f) = u_f \\ x_n^{(3)}(0) = v_n(0) \ = v_0 & x_n^{(3)}(t_f) = v_n(t_f) = v_f \\ x_n^{(4)}(0) = \lambda_{r_n}(0) = 1 & \end{array}$$

$$n = 1,2,\dots .$$

Setting $\lambda_r(0) = 1$ accomplished the scaling of the multipliers. The iteration proceeds until $\bar{\rho}(X_{n+1},X_n) \leq \beta$ where

$$\bar{\rho}(X_{n+1},X_n) = \sum_{i=1}^{6} \max_{t\in[0,t_{f_k}]} |x_{n+1}^{(i)} - x_n^{(i)}|$$

At this stage the final time, t_{f_k}, is adjusted automatically according to the difference $[r_f - r(t_{f_k})]$ by a scalar application of the Newton-Raphson procedure as follows

$$t_{f_{k+1}} = t_{f_k} + \frac{(t_{f_k} - t_{f_{k-1}})}{r(t_{f_k}) - r(t_{f_{k-1}})}[r_f - r(t_{f_k})]$$

The above iteration on X_n now continues for the new final time $t_{f_{k+1}}$ until $\bar{\rho}$ is again $\leq \beta$. The over-all process proceeds until $\rho \leq \epsilon$ where

$$\rho = \bar{\rho} + \frac{1}{b}|t_{f_{k+1}} - t_{f_k}|$$

and b is a scaling factor. The corresponding iterate X_{n+1} is accepted as the solution to the minimum time problem, and a final check is run by integrating the nonlinear Euler-Lagrange equations with a complete set of initial conditions taken from the final iterate.

For the purpose of numerical precision the data for the sample problem were normalized to obtain

$r_0 = 1.000$	$v_f = .8098$
$r_f = 1.525$	$u_f = 0.000$
$k = 1.000$	$m_0 = 1.000$
$v_0 = 1.000$	$\dot{m} = -.07487$
$u_0 = 0.000$	$T = .1405$

This resulted in a time unit of 58.18 days. The starting vector $X_0(t)$ was chosen rather crudely as follows:

t_{f_0} = 178.0 days, or 3.060 of our time units

$$x_0^{(1)}(t) = r_0(t) = r_0 + \frac{r_f - r_0}{t_{f_0}} t$$

$$x_0^{(2)}(t) = u_0(t) \equiv 0$$

$$x_0^{(3)}(t) = v_0(t) = \left(\frac{k}{r_0(t)}\right)^{\frac{1}{2}}$$

$$x_0^{(4)}(t) = \lambda_{r_0}(t) \equiv 1.000$$

$$x_0^{(5)}(t) = \lambda_{u_0}(t) \equiv \left\{ \begin{array}{lll} .5200 & \text{for} & t \epsilon [0, \frac{1}{2} t_{f_0}] \\ -.5000 & \text{for} & t \epsilon (\frac{1}{2} t_{f_0}, t_{f_0}] \end{array} \right\}$$

$$x_0^{(6)}(t) = \lambda_{v_0}(t) \equiv \left\{ \begin{array}{lll} .3000 & \text{for} & t \epsilon [0, \frac{1}{2} t_{f_0}] \\ 0.000 & \text{for} & t \epsilon (\frac{1}{2} t_{f_0}, t_{f_0}] \end{array} \right\}$$

The final two starting functions $\lambda_{u_0}(t)$ and $\lambda_{v_0}(t)$ correspond to a control angle $\theta_0(t)$ which is constant at 60° above the local horizontal for the first half of the transit time, and constant inward along the local vertical for the remaining half of the transit time (see Fig. 1).

The sequence $\{X_n\}$ converged uniformly to an accuracy of 5 significant figures with 4 shifts of the final time in 13 total iterations. The resultant minimum time was found to be 193.2 days; in agreement with results previously obtained by gradient methods (Ref. 5). The total computer time (IBM 7094) required was 36 seconds. Figure 1 illustrates the behavior of the control angle program, where $\theta_0(t)$ is the starting function, $\theta_1(t)$ through $\theta_4(t)$ correspond to the 4 shifts of the final time t_f, and $\theta^*(t)$ results from the integration of the nonlinear state and Euler-Lagrange equations with the initial values taken from the final iterate. The curves for $\theta_2(t)$, $\theta_3(t)$, and $\theta_4(t)$ lie, within our plotting accuracy, on the solution curve $\theta^*(t)$; except for the final segments as indicated on the figure. The behavior of the metric ρ is shown in Fig. 2.

We observe that for this particular example the approach just described is systematic, simple to apply, and yields rapid convergence from crude a priori starting functions.

By simple changes in the initial data, solutions were also generated for Earth to Venus and Earth to Jupiter transfers. The minimum times for these were 139.2 days and 478.2 days respectively.

CONCLUSIONS

The numerical examples of this paper suggest that the Newton-Raphson operator technique may be a useful computational method for obtaining solutions to meaningful nonlinear boundary value problems; and in particular for obtaining extremals for variational problems. It may be of particular use in generating families of solutions for given variational problems with differing values for the relevant parameters; for in this case the solution for one set of parameters becomes the starting funtion for the succeeding problem. This implies that the desired family may be generated with reasonable computation time.

We note, however, certain reservations. Although it was possible, for the included examples, to obtain crude a priori starting functions sufficient to produce convergence, it is not clear that this will remain true for other more complex problems. If it should occur that starting functions sufficient for convergence are not easily obtainable then one might consider a hybrid approach, e.g., using a few steps of a gradient technique to produce the necessary starting functions.

It should also be noted that the solutions obtained by this method are simply extremals and are not necessarily solutions of the given maximization or minimization problem. In general, further information must be brought to bear to decide whether or not one has in fact produced a solution to the optimization problem. This may be in the form of physical reasoning based upon properties of the particular system, or in the form of additional mathematical tests, e.g., the Legendre-Clebsch condition, the Weierstrass test, etc.

Finally, we observe that application of this algorithm to problems with bounded control variables and/or state variable constraints requires further modification and extension of the technique. A problem of bounded control is presently under study and will be reported upon at a later date.

ACKNOWLEDGMENT

This is to acknowledge the able assistance of Gerald E. Taylor who accomplished the necessary machine programming for all the numerical examples.

Portions of this work were generated in connection with AFOSR Contract Number AF49(638)-1207.

REFERENCES

1. Hestenes, M.R., Numerical Methods of Obtaining Solutions of Fixed End Point Problems in the Calculus of Variations, RM-102, The Rand Corporation, August 1949.

2. Kalaba, R., "On Nonlinear Differential Equations, the Maximum Operation, and Monotone Convergence," Journal of Mathematics and Mechanics, Vol. 8, No. 4, pp. 519-574, July 1959.

3. McGill, R., and P. Kenneth, "A Convergence Theorem on the Iterative Solution of Nonlinear Two-Point Boundary Value Systems," presented at the XIVth IAF Congress, Paris, France, September 1963.

4. Kolmogorov, A.N., and Fomin, S.V., Elements of the Theory of Functions and Functional Analysis, Vol. 1, Graylock Press, Rochester, New York, pp. 43 ff., 1957.

5. Kelley, H.J., "Method of Gradients," Chapter 6 of Optimization Techniques, edited by G. Leitmann, Academic Press, 1962.

TABLE I

t X	0.	0.4	0.8	1.2	1.6	2.0
x_0	1.076000	0.860800	0.645600	0.430400	0.215200	0.
x_1	1.076000	1.015153	0.845061	0.610986	0.323847	0.
x_2	1.076000	1.048799	0.900816	0.657001	0.346085	0.
x_3	1.076000	1.049839	0.902586	0.658550	0.346867	0.
x_4	1.076000	1.049840	0.902587	0.658551	0.346868	0.
x^*	1.076000	1.049840	0.902587	0.658551	0.346868	0.
y_0	0.	0.115200	0.230400	0.345600	0.460800	0.576000
y_1	0.	0.172927	0.324202	0.447591	0.537713	0.576000
y_2	0.	0.184664	0.348339	0.475158	0.553667	0.576000
y_3	0.	0.185100	0.349180	0.476056	0.554172	0.576000
y_4	0.	0.185100	0.349180	0.476057	0.554173	0.576000
y^*	C.	0.185100	0.349180	0.476057	0.554173	0.576000
z_0	0.	0.199532	0.399064	0.598597	0.798129	0.997661
z_1	0.	0.299519	0.561534	0.775250	0.931347	0.997661
z_2	0.	0.319848	0.603341	0.822998	0.958980	0.997661
z_3	0.	0.320602	0.604800	0.824553	0.959854	0.997661
z_4	0.	0.320603	0.604798	0.824555	0.959855	0.997661
z^*	0.	0.320603	0.604798	0.824555	0.959855	0.997661
X t	0.	0.4	0.8	1.2	1.6	2.0

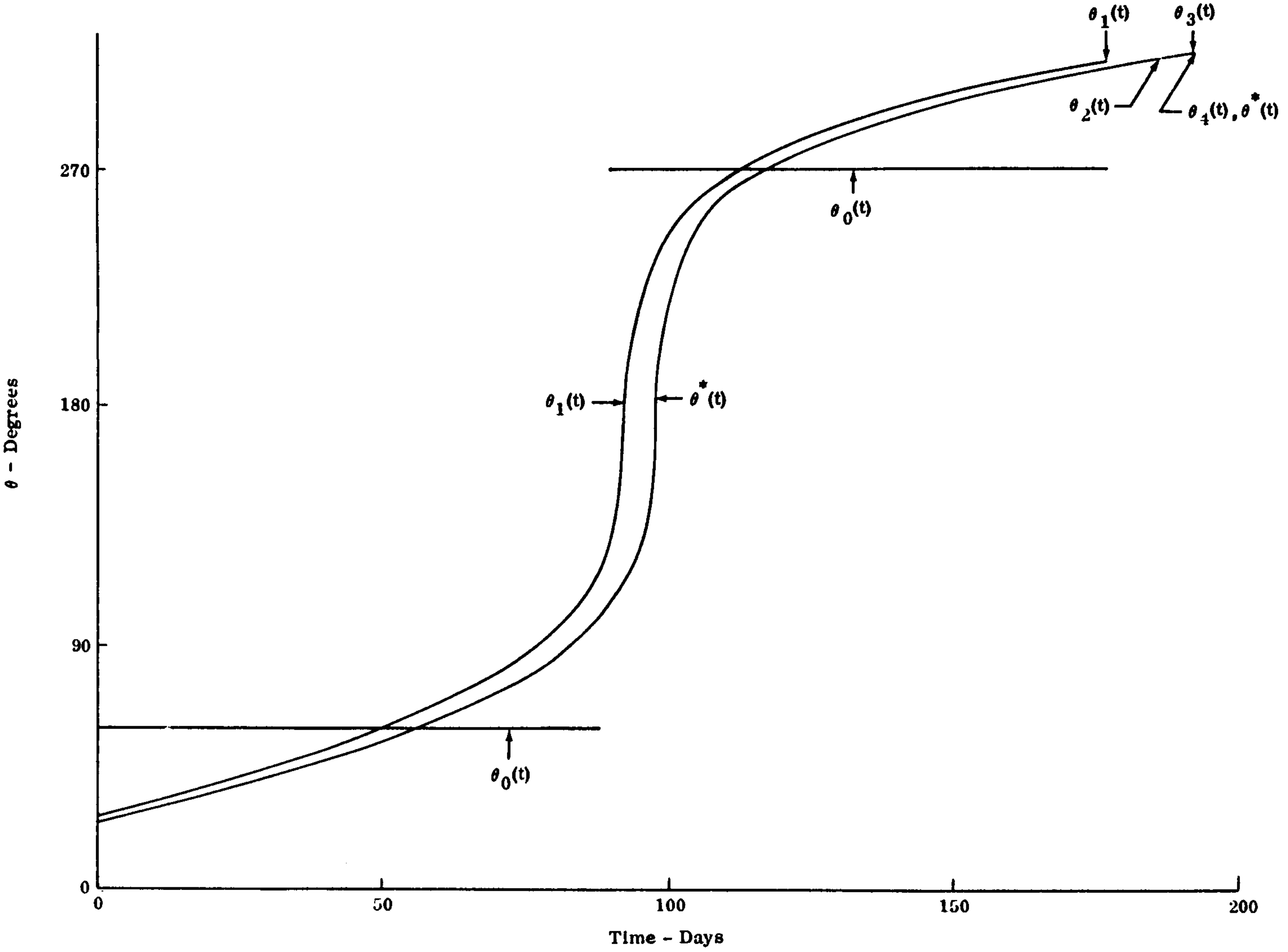


Fig. 1 Control Angle Programs from Generalized Newton-Raphson Method

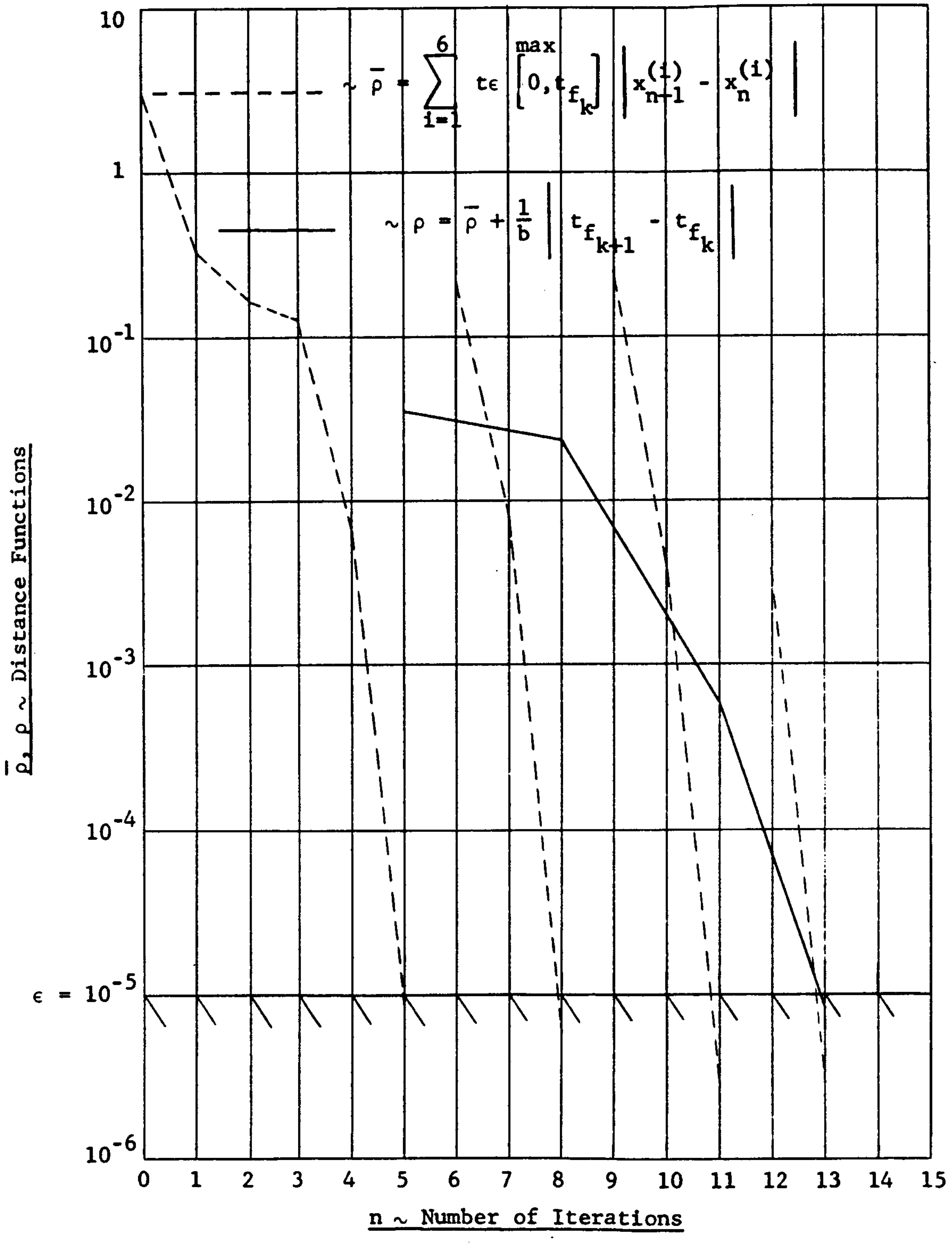


Fig. 2 $\bar{\rho}$, ρ Versus n for the Newton-Raphson Method

AUBURN UNIVERSITY

AN INVESTIGATION OF MINIMUM DRAG ATMOSPHERIC RE-ENTRY PATHS

By

Grady Harmon
W. A. Shaw

AUBURN, ALABAMA

AUBURN UNIVERSITY
AUBURN, ALABAMA

AN INVESTIGATION OF MINIMUM DRAG ATMOSPHERIC RE-ENTRY PATHS

By

Grady Harmon
W. A. Shaw

20952

A

SUMMARY

The Maximum Principle of Pontryagin is used to find the point-to-point re-entry trajectory of a space vehicle with an offset center of gravity which will minimize the accumulated aerodynamic acceleration. The mathematical model used incorporates the yaw angle of attack as the control variable and eliminates undesirable oscillations due to time variations of the rotation state variables. The set of characteristic differential equations is written with the first order equations of motion as constraints. A computation procedure is devised so that numerical solutions can be obtained on a digital computer.

Author

LIST OF SYMBOLS

G	Gravitational constant
m	Mass of the vehicle
M	Mass of the earth
$\overline{X}$	Plumbline position vector
$\overline{x}_m$	Missile system position vector
$\overline{x}_a$	Aerodynamic system position vector
$\overline{x}_{cp}$	Position of the center of pressure in the missile system
$\overline{z}_r$	Roll jet positions in the missile system
$\lvert R \rvert$	Absolute value of the plumbline position vector
R_o	Earth's radius
ϕ_p	Pitch angle
ϕ_y	Yaw angle
ϕ_r	Roll angle
SP	Sine ϕ_p
CP	Cosine ϕ_p
SY	Sine ϕ_y
CY	Cosine ϕ_y
SR	Sine ϕ_r
CR	Cosine ϕ_r
$\overline{F}_a$	Aerodynamic force in the aerodynamic coordinate system
$\overline{F}_{am}$	Aerodynamic force in the missile system
$\overline{F}_G$	Gravitational force in the plumbline system
$\overline{F}_{r_1} = - \overline{F}_{r_2}$	Roll forces in the missile system

$\bar{M}_{am}$	Aerodynamic moment (missile system)
$\bar{M}_{rm}$	Roll moment (missile system)
A	Projected cross-sectional area of vehicle
q	Dynamic pressure
$f\ (\alpha)$	Vehicle configuration function
$\bar{\omega}_E$	Earth's angular velocity vector in plumbline system
$\bar{\omega}$	Vehicle's-angular velocity vector in missile system
T	Kinetic energy
t	Time
$\bar{V}_R$	Relative velocity vector (Plumbline System)
$\bar{V}_r$	Relative velocity vector (Aerodynamic System)
$\bar{V}_{rm}$	Relative velocity vector (Missile System)
$\bar{W}$	Velocity vector for abnormal air movement in plumbline system

I. INTRODUCTION

In "Progress Report No. 4 On Studies in the Fields of Space Flight and Guidance Theory" a paper entitled "Preliminary Investigation on Six Dimensional Optimum Re-entry Trajectories" is presented by Douglas Raney and W. A. Shaw. The following paper is a continuation of that study.

In this paper the optimum re-entry problem is studied as in Progress Report No. 4 with the following exceptions:

(1) The yaw angle of attack is taken to be the single control variable.

(2) Undesirable oscillations due to time variations of the rotational state variables are eliminated.

(3) The Maximum Principle of Pontryagin is used rather than the classical calculus of variations.

II. STATEMENT OF THE PROBLEM

The problem herein presented is that of determining from a given class of allowable trajectories the best one yielding mission fulfillment.

A space vehicle is assumed to initiate a re-entry into the earth's atmosphere from some initial point above the earth's surface. The influencing forces are the gravitational force of the earth and the aerodynamic force created by atmospheric drag. The prediction of the vehicle's performance is based on the assumption that a control system is desired which will satisfy the following criteria:

1. Minimization of the accumulated g-forces on the vehicle's occupants.
2. Capability of making a point landing.

In mathematical form the first of these becomes the minimization of the integral of the square of the total aerodynamic acceleration. The second can be accomplished by the proper choice of the initial auxiliary variables.

The performance problem thus formulated becomes the fixed end point problem of Lagrange, where the integral to be minimized has as constraints the first order equations of motion of the vehicle. The boundary conditions are the initial and terminal values of position, velocity, and roll angle. The magnitude of the yaw angle of attack is taken as the control variable.

Additional assumptions made are as follows:

1. The earth is a rotating sphere and the inverse gravity law holds.
2. The mass of the vehicle is invariant with respect to time.
3. The vehicle has an offset center of gravity which is invariant with respect to the vehicle.
4. A pure couple is produced about the roll axis of the vehicle by properly placed jets whose force magnitudes are functions of the control variable.
5. The angular velocity and the angular acceleration of the pitch and yaw angles are zero and the acceleration of the roll angle is zero.
6. The center of pressure is invariant with respect to the center of gravity.

III. COORDINATE SYSTEMS

Three rectangular cartesian coordinate systems will be used in this paper. They are:

1. The plumbline space fixed coordinate system
2. The vehicle fixed missile system
3. The aerodynamic system.

A. PLUMBLINE SYSTEM

The plumbline system, Figure 1, has its origin at the earth's center with the Y axis parallel to the gravity gradient at the launch point. The X axis is parallel to the earth fixed launch azimuth and the Z axis is such as to form a right-handed system.

B. MISSILE SYSTEM

The missile system, Figure 1, is defined with its origin at the center of gravity of the vehicle and its y_m axis parallel to the longitudinal axis of the vehicle. The x_m and z_m axes are taken so as to form a right-handed system which is parallel to the plumbline system at the launch point.

As the vehicle moves along its trajectory, the missile system undergoes a displacement with respect to the plumbline system. In flight the two coordinate systems are related through Eulerian angles which are measured by a gimbal. The direction of the vehicle in flight is defined by first rotating about the Z axis by ϕ_p, then around the new intermediate x axis by ϕ_y, and finally around the

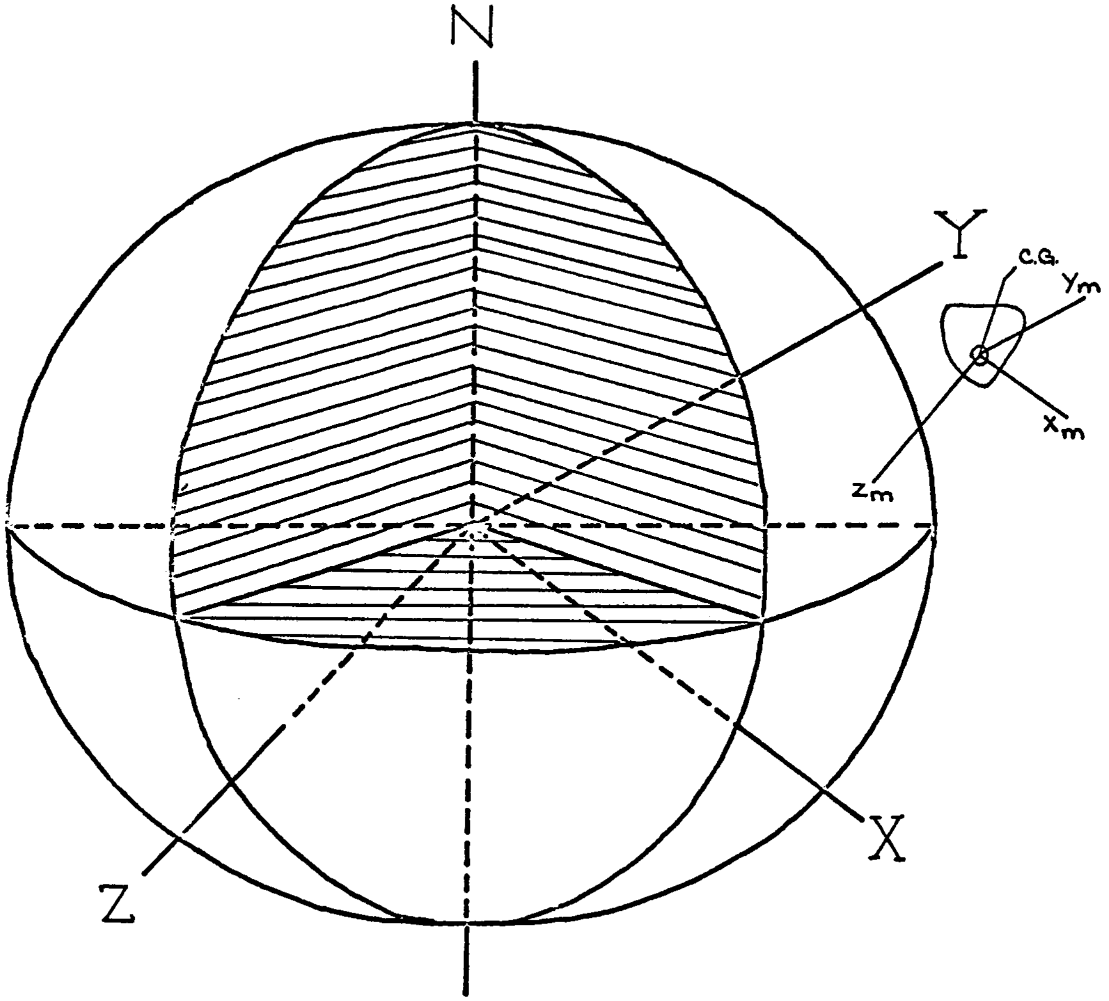


FIGURE 1.

PLUMBLINE AND MISSILE COORDINATE SYSTEMS

y_m axis by $-\phi_r$. Thus, a position vector in the missile system may be written in terms of the position vector in the plumbline system as

$$\bar{x}_m = \left[-\phi_r\right]_2 \left[\phi_y\right]_1 \left[\phi_p\right]_3 \bar{X} \quad , \tag{1}$$

or

$$\begin{bmatrix} x_m \\ y_m \\ z_m \end{bmatrix} = \begin{bmatrix} CR & 0 & SR \\ 0 & 1 & 0 \\ -SR & 0 & CR \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & CY & SY \\ 0 & -SY & CY \end{bmatrix} \begin{bmatrix} CP & SP & 0 \\ -SP & CP & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} . \tag{1a}$$

Expanding the above gives

$$\bar{x}_m = \left[\begin{array}{c:c:c} CRCP + SRSYSP & CRSP - SRSYCP & SRCY \\ -CYSP & CYCP & SY \\ -SRCP + CRSYSP & -SPSR - CRCPSY & CRCY \end{array}\right] \bar{X} = \left[A_D\right] \bar{X}, \tag{1b}$$

where $\left[A_D\right]$ is the transformation matrix and CR, for example, is used to denote cosine ϕ_r. The gimbal angles are illustrated in Figure 2 where a right hand rotation is positive. The above definitions of Eulerian angles are consistent with those used in computer decks compiled by NASA.(4)

AERODYNAMIC SYSTEM

The aerodynamic system is defined with its origin at the center of pressure of the vehicle and its y_a axis coincident with the relative velocity vector. The x_a and z_a axes are chosen to form a right hand system.

Again, as the vehicle moves in flight, there will be a displacement of the missile and aerodynamic coordinate systems relative to one another. The direction of the relative velocity vector or the y_a axis may be defined by the following rotations:

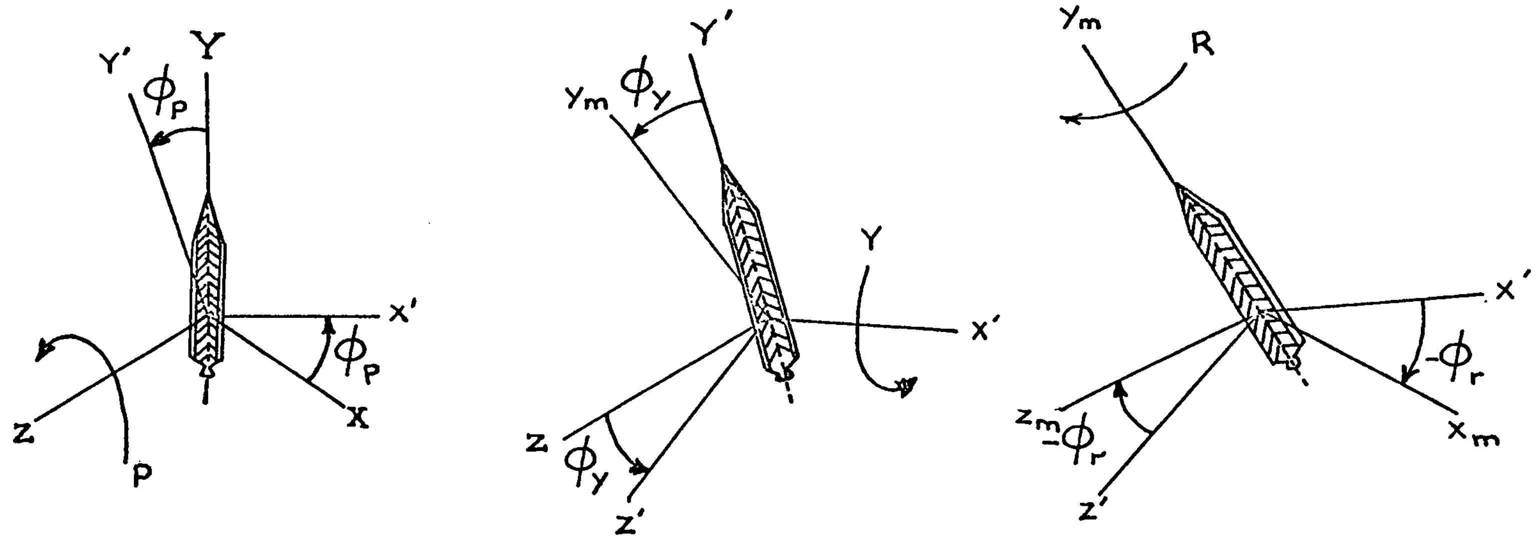


FIGURE 2. EULERIAN ANGLES

1. Rotate the vehicle fixed reference frame about the y_m axis such that the x_m axis is brought to lie in the plane which contains the y_m axis and the relative velocity vector. Denote this angle as α_y.

2. Rotate about the new z axis to bring the y_m axis coincident with the relative velocity vector. Denote this angle as α . This angle is the so-called true angle of attack.

A position vector may now be written in the aerodynamic system in terms of the missile system as

$$\bar{x}_a = \left[-\alpha\right]_3 \left[\alpha_y\right]_2 \bar{x}_m \quad , \tag{2}$$

or

$$\begin{bmatrix} x_a \\ y_a \\ z_a \end{bmatrix} = \begin{bmatrix} c\alpha & -s\alpha & 0 \\ s\alpha & c\alpha & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} c\alpha_y & 0 & -s\alpha_y \\ 0 & 1 & 0 \\ s\alpha_y & 0 & c\alpha_y \end{bmatrix} \begin{bmatrix} x_m \\ y_m \\ z_m \end{bmatrix} . \tag{2a}$$

$$\bar{x}_a = \left[\begin{array}{c:c:c} c\alpha \; c\alpha_y & -s\alpha & -c\alpha \; s\alpha_y \\ c\alpha_y \; s\alpha & c\alpha & -s\alpha \; s\alpha_y \\ s\alpha_y & 0 & c\alpha_y \end{array}\right] \bar{x}_m = \left[A_a\right] \bar{x}_m . \tag{2b}$$

Figure 3 illustrates this system.

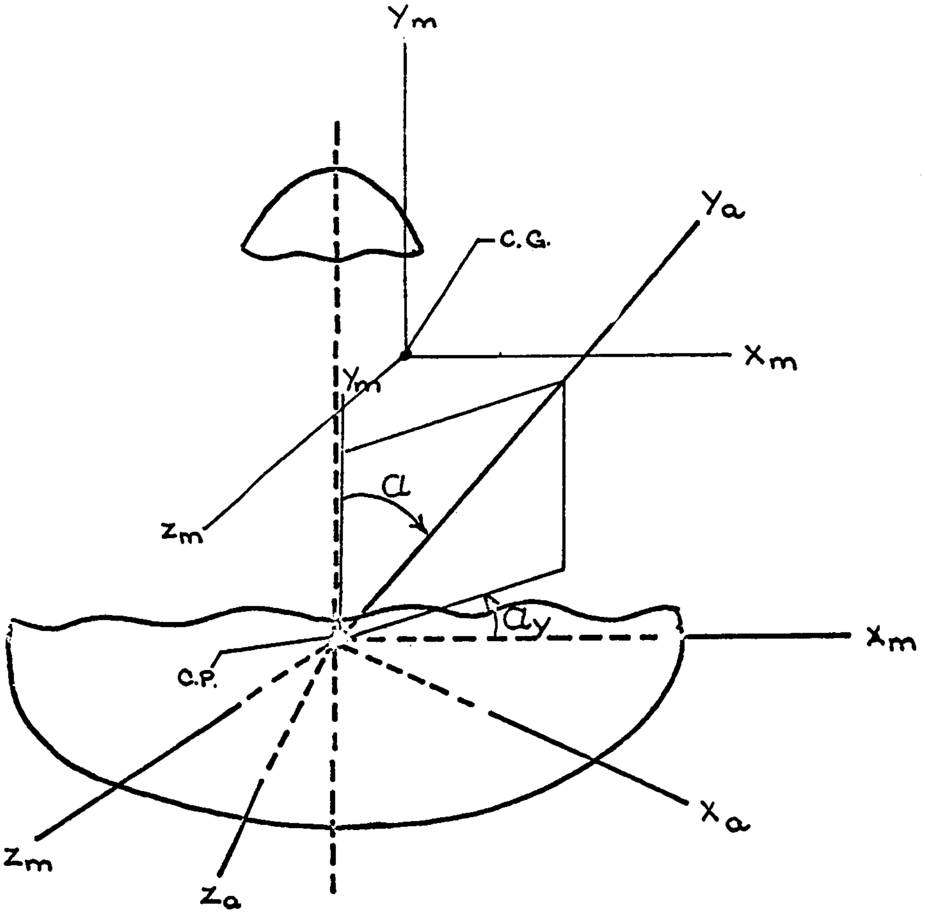


FIGURE 3. MISSILE AND AERODYNAMIC COORDINATE SYSTEMS

IV. BASIC MECHANICS

A. FORCES

Gravitational force. Since a spherical earth was assumed, Newton's Law of Universal Gravitation which gives us an attractive force between the earth and the vehicle is

$$\bar{F}_G = - \frac{GMm\bar{X}}{|R|^3} \tag{3}$$

Aerodynamic force. The aerodynamic force, Figure 4, is a force due to atmospheric drag. It acts through the center of pressure and the direction of the force is always parallel and opposite to the relative velocity vector. Written in the aerodynamic system the force takes the following form:

$$\bar{F}_a = \begin{bmatrix} 0 \\ -F_a \\ 0 \end{bmatrix} . \tag{4}$$

In the missile system

$$\bar{F}_{am} = \left[A_a\right]^T \bar{F}_a \quad , \tag{5}$$

or

$$\bar{F}_{am} = \begin{bmatrix} F_{amx} \\ F_{amy} \\ F_{amz} \end{bmatrix} = \begin{bmatrix} -F_a & s\alpha & c\alpha_y \\ -F_a & c\alpha & \\ F_a & s\alpha & s\alpha_y \end{bmatrix} . \tag{5a}$$

The expression for the magnitude of F_a is taken to be the same as that proposed by Miner(4), i.e., $F_a = Aqf(\alpha)$. A is the projected cross-section area of the vehicle, q the dynamic pressure, and $f(\alpha)$ a factor which is determined by the vehicle's configuration.

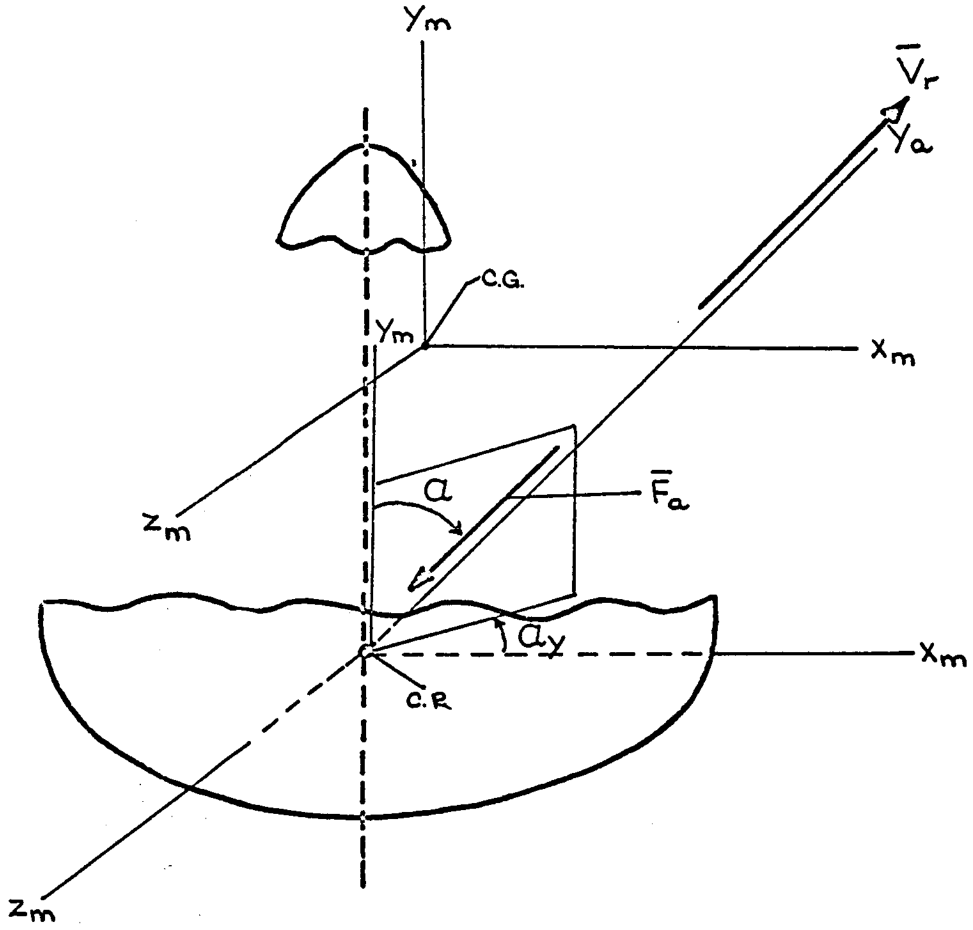


FIGURE 4. AERODYNAMIC FORCE SYSTEM

Since the aerodynamic force is dependent upon the relative velocity or the flow of air over the missile, it is appropriate at this time to discuss this flow. Miner's proposals are again used where he assumes that the atmosphere in the large moves with the earth. This gives at all times an air mass movement with respect to the plumbline system of

$$\bar{X} \times \bar{\omega}_E - \bar{W} \quad ,$$

where $\bar{W}$ is used to represent any abnormal air movement desired. The relative velocity vector in the plumbline system is then given by

$$\bar{V}_R = \dot{\bar{X}} + \left[\bar{X} \times \bar{\omega}_E - \bar{W}\right] \quad , \qquad (6)$$

or

$$\begin{bmatrix} V_{RX} \\ V_{RY} \\ V_{RZ} \end{bmatrix} = \begin{bmatrix} \dot{X} \\ \dot{Y} \\ \dot{Z} \end{bmatrix} + \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} \times \begin{bmatrix} \omega_{EX} \\ \omega_{EY} \\ \omega_{EZ} \end{bmatrix} - \begin{bmatrix} W_X \\ W_Y \\ W_Z \end{bmatrix} \quad . \qquad (6a)$$

In the missile system the relative velocity may be written as

$$\bar{V}_{rm} = \left[A_D\right] \bar{V}_R = \begin{bmatrix} V_{rmx} \\ V_{rmy} \\ V_{rmz} \end{bmatrix} \quad , \qquad (7)$$

or in terms of the aerodynamic system variables

$$\bar{V}_{rm} = \left[A_a\right]^T \bar{V}_r \quad , \qquad (8)$$

where

$$\bar{V}_r = \begin{bmatrix} 0 \\ V_r \\ 0 \end{bmatrix} \quad .$$

B. MOMENTS

Aerodynamic moment. Since both the center of pressure and the center of gravity are invariant with respect to the vehicle, a constant vector in the missile system may be used to relate the two. Let $\bar{x}_{cp}$ be the missile fixed coordinates of the center of pressure with respect to the center of gravity. The aerodynamic moment about the center of gravity is then given by

$$\bar{M}_{am} = \bar{x}_{cp} \times \bar{F}_{am} \quad , \tag{9}$$

or

$$\begin{bmatrix} M_{amx} \\ M_{amy} \\ M_{amz} \end{bmatrix} = \begin{bmatrix} y_{cp} F_a \, s\alpha \, s\alpha_y + z_{cp} F_a \, c\alpha \\ -x_{cp} F_a \, s\alpha \, s\alpha_y - z_{cp} F_a \, s\alpha \, c\alpha_y \\ -x_{cp} F_a \, c\alpha + y_{cp} F_a \, s\alpha \, c\alpha_y \end{bmatrix} . \tag{9a}$$

Roll moment. Reference to Figure 5 will show the system of roll jets which is used to fulfill assumption 4 of the problem statement. The jets are placed so that in the missile system

$$\bar{F}_{r_1} = \begin{bmatrix} F_r \\ 0 \\ 0 \end{bmatrix} , \text{ located at } \bar{z}_r = \begin{bmatrix} 0 \\ 0 \\ z_r \end{bmatrix} ,$$

and $$\bar{F}_{r_2} = \begin{bmatrix} -F_r \\ 0 \\ 0 \end{bmatrix} , \text{ located at } -\bar{z}_r = \begin{bmatrix} 0 \\ 0 \\ -z_r \end{bmatrix} .$$

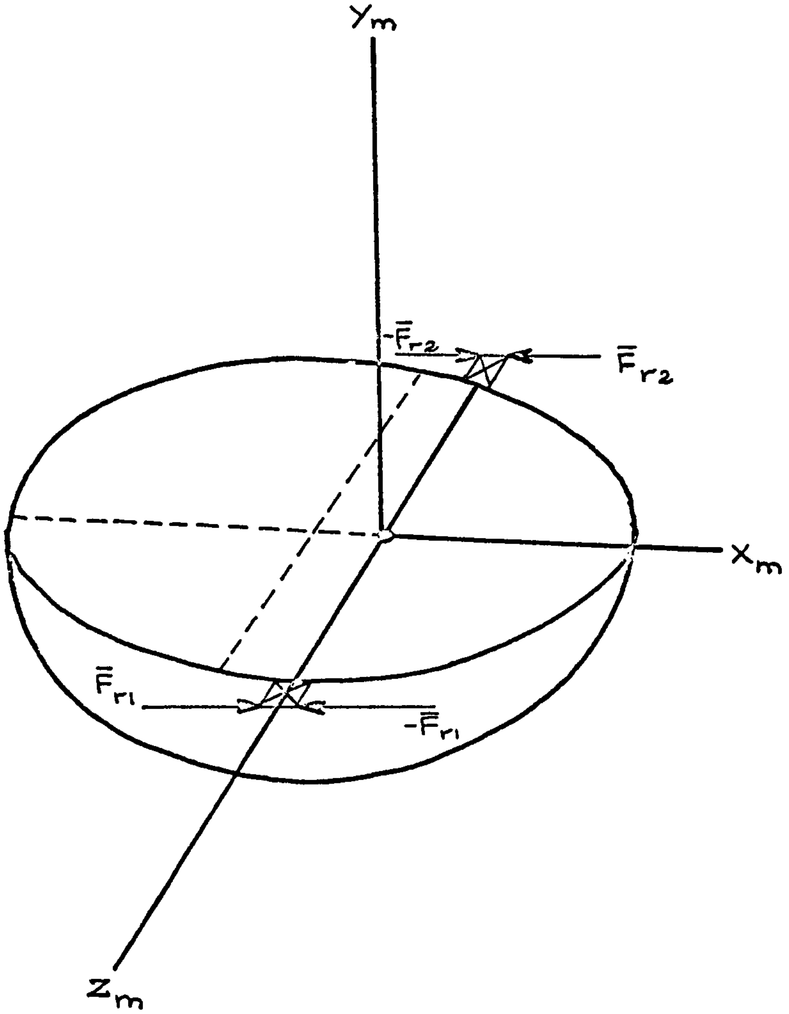


FIGURE 5.

ROLL FORCE SYSTEM

The moment about the center of gravity caused by these forces is thus given by

$$\bar{M}_{rm} = 2\left[\bar{z}_r \times \bar{F}_{r_1}\right] , \tag{10}$$

since $\bar{F}_{r_1} = -\bar{F}_{r_2}$,

or

$$\bar{M}_{rm} = \begin{bmatrix} 0 \\ 2z_r F_r \\ 0 \end{bmatrix} . \tag{10a}$$

The total moment about the center of gravity, in the missile system, is then the sum of the aerodynamic and roll moments.

$$\bar{M}_{T_m} = \bar{M}_{am} + \bar{M}_{rm} , \tag{11}$$

or

$$\bar{M}_{T_m} = \begin{bmatrix} y_{cp}\ F_a\ s\alpha\, s\alpha_y + z_{cp}\ F_a\ c\alpha \\ -x_{cp}\ F_a\ s\alpha\, s\alpha_y - z_{cp}\ F_a\ s\alpha\, c\alpha_y + 2F_r z_r \\ -x_{cp}\ F_a\ c\alpha + y_{cp}\ F_a\ s\alpha\, c\alpha_y \end{bmatrix} . \tag{11a}$$

V. EQUATIONS OF MOTION

From Chasle's theorem of mechanics, it is possible to interpret the equations of motion of a rigid body as the sum of two independent effects. One, the motion of the center of gravity with respect to the inertial coordinate system and two, the motion of the rigid body around its center of gravity. In general, this type of rigid body motion in three-dimensional space requires six degrees of freedom since six state variables are needed to fix the orientation of the body with respect to the inertial frame. The state variables used in this problem are the plumbline coordinates and the Eulerian angles.

A. TRANSLATION MOTION

As previously stated, only gravitational and aerodynamic forces are considered. Using Newton's Second Law, the translational motion of the center of gravity with respect to the plumbline system is given by the following set of second order differential equations.

$$\ddot{\overline{X}} = - \frac{GM\overline{X}}{|R|^3} + \frac{[A_D]^T}{m} \overline{F}_{am} , \tag{12}$$

where

$$\ddot{\overline{X}} = \begin{bmatrix} \ddot{X} \\ \ddot{Y} \\ \ddot{Z} \end{bmatrix}$$

By making the following change of variable, the second order equations of translational motion may be reduced to first order.

$$\bar{u} = \begin{bmatrix} u \\ v \\ w \end{bmatrix} \equiv \begin{bmatrix} \dot{X} \\ \dot{Y} \\ \dot{Z} \end{bmatrix} = \dot{\bar{X}} \qquad . \tag{13}$$

The first order translational equations thus become

$$\dot{\bar{u}} = - \frac{GM\bar{X}}{|R|^3} + \frac{[A_D]^T}{m} \bar{F}_{am} \qquad . \tag{14}$$

For convenience, the following definitions are made:

$$g \equiv - \frac{GM}{|R|^3} \quad , \tag{15}$$

$$\frac{[A_D]^T}{m} \bar{F}_{am} \equiv \frac{F_a}{m} \bar{N} \equiv F_a' \bar{N} \quad , \tag{16}$$

where

$$\bar{N} = \begin{bmatrix} N \\ P \\ Q \end{bmatrix} = \begin{bmatrix} -(CRCP + SRSPSY)(S\alpha C\alpha_y) + CYSPC\alpha + (-SRCP + CRSYSP)(S\alpha S\alpha_y) \\ -(CRSP - SYCPSR)(S\alpha C\alpha_y) - CYCPC\alpha - (SPSR + CRCPSY)(S\alpha S\alpha_y) \\ -(SRCY)(S\alpha C\alpha_y) - SYC\alpha + (CRCY)(S\alpha S\alpha_y) \end{bmatrix} . \tag{17}$$

Thus, the translational equations may be written as

$$\dot{\bar{u}} = F_a' \bar{N} + g \bar{X} \, . \tag{18}$$

B. ROTATIONAL MOTION

In writing the rotational equations of motion the energy method or the Lagrangian form was found to be more convenient than the Newtonian approach. For generalized coordinates of angular character, such as the Eulerian set, the Lagrangian form becomes

$$\frac{d}{dt}\left(\frac{\partial T}{\partial \dot{\phi}_i}\right) - \frac{\partial T}{\partial \phi_i} = M\phi_i \;, \tag{19}$$

$$i = p, y, r,$$

where $M\phi_i$ is the moment associated with the ϕ_i rotation and T is the kinetic energy. Since an offset center of gravity was assumed, all components of the inertia matrix are taken to be non-zero. The kinetic energy for such a system is given by

$$T = \tfrac{1}{2}\, \bar{\omega}^T [\mu]\, \bar{\omega} \;, \tag{20}$$

where $[\mu]$ is the inertia matrix

$$[\mu] = \begin{bmatrix} I_{xx} & -I_{xy} & -I_{xz} \\ -I_{xy} & I_{yy} & -I_{yz} \\ -I_{xz} & -I_{yz} & I_{zz} \end{bmatrix} \;, \tag{21}$$

and $\bar{\omega}$ is the angular velocity vector of the vehicle written in the missile fixed system. Using the expression above for the kinetic energy, the Lagrangian equation takes the following form:

$$\frac{d}{dt}\left(\frac{\partial \bar{\omega}^T}{\partial \dot{\phi}_i}\right)[\mu]\,\bar{\omega} + \frac{\partial \bar{\omega}^T}{\partial \phi_i}[\mu]\,\frac{d\bar{\omega}}{dt} - \frac{\partial \bar{\omega}^T}{\partial \phi_i}[\mu]\,\bar{\omega} = M\phi_i . \tag{22}$$

The $\bar{\omega}$ vector is obtained by transforming the angular velocity components $\dot{\phi}_p$, $\dot{\phi}_y$, and $\dot{\phi}_r$ into the missile system from their positions in the directions of the axes of rotation. Since the gimbal system used in this analysis measures pitch, yaw, and roll in that order, turning from the space fixed plumbline system, the following transformations must be made. $\dot{\phi}_r$ is already in the missile system; $\dot{\phi}_y$

must be rotated through ϕ_r, and $\dot{\phi}_p$ must be rotated through ϕ_y and then ϕ_r. In the missile system, the $\bar{\omega}$ vector thus becomes

$$\bar{\omega} = \begin{bmatrix} 0 \\ \dot{\phi}_r \\ 0 \end{bmatrix} + \begin{bmatrix} CR & 0 & SR \\ 0 & 1 & 0 \\ -SR & 0 & CR \end{bmatrix} \begin{bmatrix} \dot{\phi}_y \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} CR & 0 & SR \\ 0 & 1 & 0 \\ -SR & 0 & CR \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & CY & SY \\ 0 & -SY & CY \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ \dot{\phi}_p \end{bmatrix} \qquad (23)$$

or

$$\bar{\omega} = \begin{bmatrix} SRCY & CR & 0 \\ SY & 0 & -1 \\ CRCY & -SR & 0 \end{bmatrix} \begin{bmatrix} \dot{\phi}_p \\ \dot{\phi}_y \\ \dot{\phi}_r \end{bmatrix} = \left[A_\omega\right] \dot{\bar{\phi}} \quad . \qquad (23a)$$

Also

$$\bar{\omega}^T = \dot{\bar{\phi}}^T \left[A_\omega\right]^T \quad .$$

Using these expressions in the rotational equations and rewriting in vector form, the Lagrange equations in pitch, yaw, and roll become:

$$\ddot{\bar{\phi}} = [C] \left\{ \bar{M}_\phi - [S]\dot{\bar{\phi}} - [T]\dot{\bar{\phi}} + \bar{B} \right\} , \qquad (24)$$

where $B_i = \dot{\bar{\phi}}^T \dfrac{\partial \left[A_\omega\right]^T}{\partial \phi_i} \left[\mu\right] \left[A_\omega\right] \dot{\bar{\phi}}$,

and

$$\bar{B} = \begin{bmatrix} B_p \\ B_y \\ B_r \end{bmatrix} , \quad \ddot{\bar{\phi}} = \begin{bmatrix} \ddot{\phi}_p \\ \ddot{\phi}_y \\ \ddot{\phi}_r \end{bmatrix} , \quad \dot{\bar{\phi}} = \begin{bmatrix} \dot{\phi}_p \\ \dot{\phi}_y \\ \dot{\phi}_r \end{bmatrix} ,$$

$$\bar{M}_\phi = \left[A_\omega\right]^T \bar{M}_{Tm} ,$$

$$[C] = \left\{ \left[A_\omega\right]^T \left[\mu\right] \left[A_\omega\right] \right\}^{-1} ,$$

$$\left[S\right] = \frac{d\left[A_\omega\right]^T}{dt}\left[\mu\right]\left[A_\omega\right], \text{ and}$$

$$\left[T\right] = \left[A_\omega\right]^T\left[\mu\right]\frac{d\left[A_\omega\right]}{dt} .$$

Now in order to comply with assumption five of the problem statement, the following definitions will be used throughout the remainder of this paper.

$$\ddot{\bar{\phi}} \equiv \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} = 0 \qquad (25) \qquad \dot{\bar{\phi}} \equiv \begin{bmatrix} 0 \\ 0 \\ \dot{\phi}_r \end{bmatrix} \qquad (26)$$

The first order rotational equations thus become

$$\bar{M}_{Tm} = \left\{\left[A_\omega\right]^T\right\}^{-1}\left\{\left[S\right]\dot{\bar{\phi}} + \left[T\right]\dot{\bar{\phi}} - \bar{B}\right\} , \qquad (27)$$

where it is shown in appendix one that $\left[C\right]$ is a non-zero matrix.

VI. FORMULATION OF THE VARIATIONAL PROBLEM

Before going into the mechanics of the variational problem, further consideration must be given to the equations defining relative velocity and to the constraint equations. These equations must be solved for a particular set of variables so that a computational procedure can be devised.

First to be considered will be the equations defining relative velocity. Written in two different sets of variables, the relative velocity in the missile system takes the form

$$\overline{V}_{rm} = \left[A_D\right] \overline{V}_R = \left[A_a\right]^T \overline{V}_r \quad . \qquad (7 - 8)$$

The components of this vector yield the three equations (28) through (30):

$$(CRCP + SRSYSP)\ V_{RX} + (CRSP - SRSYCP)\ V_{RY} + SRCY\ V_{RZ} = V_r\ S\alpha\ C\alpha_y \qquad (28)$$

$$-CYSP\ V_{RX} + CYCP\ V_{RY} + SY\ V_{RZ} = V_r\ C\alpha \qquad (29)$$

$$(-SRCP + CRSYSP)\ V_{RX} - (SRSP + CRCPSY)\ V_{RY} + CRCY\ V_{RZ} = -V_r\ S\alpha\ S\alpha_y . \qquad (30)$$

This set of equations is not an independent set and thus cannot be solved for three unknowns. A clue as to the dependency may be gotten by realizing that the three equations are components of a vector and that only two angles are necessary for locating a vector in three space. In order to solve the problem, as stated in this analysis, the relative velocity equations are used to obtain the two variables ϕ_p and ϕ_y.

From a combination of equations (28) and (30), the expression for ϕ_p is found.

$$SP = \frac{J\,V_{RY} - \sqrt{J^2\,V_{RY}^2 - (V_{RX}^2 + V_{RY}^2)(J^2 - V_{RX}^2)}}{(V_{RX}^2 + V_{RY}^2)} \, , \tag{31}$$

and

$$CP = \frac{J\,V_{RX} + \sqrt{J^2\,V^2{}_{RX} - (V_{RX}^2 + V_{RY}^2)(J^2 - V_{RY}^2)}}{(V_{RX}^2 + V_{RY}^2)} \, , \tag{32}$$

where $J = CR\ V_{rmx} - SR\ V_{rmz}$. $\quad$ (33)

Thus $\phi_p = \text{ARC TAN}\left(\frac{SP}{CP}\right)$; $\quad -\pi \leq \phi_p \leq \pi$. $\quad$ (34)

The solution thus obtained is not unique from a purely mathematical view; however, if physical considerations which lead to consistency in the problem are granted, then the solution is unique. (See Appendix II). Equation (29) is solved for ϕ_y.

$$SY = \frac{V_{rmy}\,V_{RZ} - \sqrt{V_{rmy}^2\,V_{RZ}^2 - (V_{RZ}^2 + K^2)(V_{rmy}^2 - K^2)}}{(V_{RZ}^2 + K^2)} \tag{35}$$

and

$$CY = \frac{V_{rmy}\,K + \sqrt{V_{rmy}^2\,K^2 - (V_{RZ}^2 + K^2)(V_{rmy}^2 - V_{RZ}^2)}}{(V_{RZ}^2 + K^2)} \, , \tag{36}$$

where

$K = CP\ V_{RY} - SP\ V_{RX}$. $\quad$ (37)

Thus the expression for ϕ_y is

$$\phi_y = \text{ARC TAN}\left(\frac{SY}{CY}\right) ; \qquad -\pi \leq \phi_y \leq \pi \, . \tag{38}$$

Again, as in solving for ϕ_p , the uniqueness of the solution comes from physical considerations. (See Appendix III.)

The components of the vector equation for rotational motion are now written and they are solved for the variables F_r, $\dot{\phi}_r$, and α . The choice of variables to be solved for is again made with the computational procedure in mind. From the vector equation (27) the components take the following form:

$$-I_{yz}\,\dot{\phi}_r^{2} = y_{cp}\,F_a\,s\alpha\,s\alpha_y + z_{cp}\,F_a\,c\alpha \qquad (39)$$

$$0 = -x_{cp}\,F_a\,s\alpha\,s\alpha_y - z_{cp}\,F_a\,s\alpha\,c\alpha_y + 2\,F_r\,z_r \qquad (40)$$

$$I_{xy}\,\dot{\phi}_r^{2} = -x_{cp}\,F_a\,c\alpha + y_{cp}\,F_a\,s\alpha\,c\alpha_y\ . \qquad (41)$$

Solving the three independent equations for three unknowns yields

$$\alpha = \text{ARC TAN}\left[\frac{I_{yz}\,x_{cp} - I_{xy}\,z_{cp}}{y_{cp}\,(I_{yz}\,c\alpha_y + I_{xy}\,s\alpha_y)}\right], \qquad (42)$$

$$F_r = \frac{F_a\,s\alpha\,(x_{cp}\,s\alpha_y + z_{cp}\,c\alpha_y)}{2\,z_r}, \qquad (43)$$

and

$$\dot{\phi}_r = \pm\sqrt{\frac{F_a\,(y_{cp}\,s\alpha\,c\alpha_y - x_{cp}\,c\alpha)}{I_{xy}}}\ . \qquad (44)$$

Equations (34), (38), (42), (43), and (44) are thus the equations which, along with the characteristic equations, form the problem solution.

As expressed in the problem statement, it is desired to determine from a given class of allowable trajectories the best one yielding mission fulfillment. This is accomplished by finding among all admissible controls $\alpha_y(t)$ which transfer the vehicle from $\bar{X}_o$ to $\bar{X}_T$ one for which the functional:

$$D = \int_{t_o}^{t_T} \left[\text{DRAG} \right]^2 dt \tag{45}$$

takes on a minimum value. In this analysis the word drag will be used synonymously with aerodynamic acceleration. Thus from Equation (18),

$$\left[\text{DRAG} \right]^2 = F_a' \, \bar{N} \cdot F_a' \, \bar{N} = (F_a')^2 \, \bar{N} \cdot \bar{N} = (F_a')^2 , \tag{46}$$

and

$$D = \int_{t_o}^{t_T} (F_a')^2 \, dt \quad (45); \qquad \dot{D} = (F_a')^2 \quad . \tag{47}$$

The Pontryagin H function may now be written as follows:

$$H = \bar{\lambda}_I \cdot \dot{\bar{X}} + \bar{\lambda}_{II} \cdot \dot{\bar{u}} + \lambda_7 \, \dot{\phi}_r + \lambda_8 \, \dot{D} \quad , \tag{48}$$

where

$$\bar{\lambda}_I = \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{bmatrix} \quad \text{and} \quad \bar{\lambda}_{II} = \begin{bmatrix} \lambda_4 \\ \lambda_5 \\ \lambda_6 \end{bmatrix} \quad .$$

The $\lambda_i(t)$, $i = 1 \ldots 8$, are the auxiliary variables that are incorporated in the same manner as the Lagrange multipliers in the classical calculus of variations. Substituting into H from Equations (18), (44), and (47) results in the following:

$$H = \bar{\lambda}_I \cdot \mathring{\bar{X}} + \bar{\lambda}_{II} \cdot \left[F_a' \, \bar{N} + g \, \bar{X} \right]$$

$$\pm \lambda_7 \sqrt{F_a} \sqrt{\frac{y_{cp} \; s\alpha \; c\alpha_y - x_{cp} \; c\alpha}{I_{xy}}} + \lambda_8 (F_a')^2 . \tag{49}$$

The expressions for the auxiliary variables are obtained from the H function and take the following form:

$$-\dot{\bar{\lambda}}_{I} = \frac{\partial H}{\partial \bar{X}} = F_a' \frac{\partial(\bar{\lambda}_{II} \cdot \bar{N})}{\partial \bar{X}} + (\bar{\lambda}_{II} \cdot \bar{N}) \frac{\partial F_a'}{\partial \bar{X}}$$

$$+ \bar{\lambda}_{II}\, g + (\bar{\lambda}_{II} \cdot \bar{X}) \frac{\partial g}{\partial \bar{X}}$$

$$\pm \lambda_7 \sqrt{\frac{y_{cp} \; s\alpha \; c\alpha_y - x_{cp} \; c\alpha}{I_{xy}}} \; \frac{\partial (F_a)^{\frac{1}{2}}}{\partial \bar{X}}$$

$$+ \lambda_8 \frac{\partial (F_a')^2}{\partial \bar{X}} \tag{50}$$

$$-\dot{\bar{\lambda}}_{II} = \frac{\partial H}{\partial \bar{u}} = \bar{\lambda}_{I} + F_a' \frac{\partial(\bar{\lambda}_{II} \cdot \bar{N})}{\partial \bar{u}} + (\bar{\lambda}_{II} \cdot \bar{N}) \frac{\partial F_a'}{\partial \bar{u}}$$

$$\pm \lambda_7 \sqrt{\frac{y_{cp} \; s\alpha \; c\alpha_y - x_{cp} \; c\alpha}{I_{xy}}} \; \frac{\partial (F_a)^{\frac{1}{2}}}{\partial \bar{u}}$$

$$+ \lambda_8 \frac{\partial (F_a')^2}{\partial \bar{u}} \tag{51}$$

$$-\dot{\lambda}_7 = \frac{\partial H}{\partial \phi_r} = F_a' \left[\bar{\lambda}_{II} \cdot \frac{\partial \bar{N}}{\partial \phi_r} \right] \tag{52}$$

$$-\dot{\lambda}_8 = \frac{\partial H}{\partial D} = 0 \tag{53}$$

It is implied from equation (53) that λ_8 = constant. The equation to be solved for the control variable and a necessary condition for a minimum of D are given below.

$$\frac{\partial H}{\partial \alpha_y} = F_a' \left(\lambda_{II} \cdot \frac{\partial \bar{N}}{\partial \alpha_y} \right)$$

$$\pm \lambda_7 \sqrt{F_a} \quad \frac{\partial}{\partial \alpha_y} \sqrt{\frac{y_{cp} \quad s\alpha \quad c\alpha_y - x_{cp} \, c\alpha}{I_{xy}}} = 0 \quad (54)$$

For a minimum of D,

$$\frac{\partial^2 H}{\partial \alpha_y^2} = F_a' \left[\lambda_{II} \cdot \frac{\partial^2 \bar{N}}{\partial \alpha_y^2} \right]$$

$$\pm \lambda_7 \sqrt{F_a} \quad \frac{\partial^2}{\partial \alpha_y^2} \sqrt{\frac{y_{cp} \quad s\alpha \, c\alpha_y - x_{cp} \quad c\alpha}{I_{xy}}} > 0 \quad (55)$$

As shown above in the Pontryagin formulation, λ_8 = constant. This constant will be chosen as $\lambda_8 = +1$ in order that a minimization of the H function will also be a minimization of D, i.e., $\frac{\partial^2 H}{\partial \alpha_y^2} > 0$ for a minimum D.

Equations (34), (38), and (42) - (44), are the constraint and definition equations which must be satisfied, and equations (50) through (54) are the characteristic equations. The complete set of algebraic and differential equations needed for the problem solution have thus been found. The desired minimum drag re-entry path will thus be one which satisfies all the aforementioned equations. A closed form solution to this set of equations does not seem probable nor is the time spent in searching for such a solution justifiable since numerical solutions via digital computers can be achieved to almost any degree of accuracy.

VII. COMPUTATIONAL SCHEME

Before the computational procedure is written, it is found convenient to rewrite important equations in functional form. Reference to these equations will be made throughout the computational scheme.

$$\alpha = \alpha(\alpha_y) \tag{56}$$

$$F_r = F_r(\bar{X}, \dot{\bar{X}}, \alpha, \alpha_y) \tag{57}$$

$$\dot{\phi}_r = \pm \dot{\phi}_r(\bar{X}, \dot{\bar{X}}, \alpha, \alpha_y) \tag{58}$$

$$\phi_p = \phi_p(\bar{X}, \dot{\bar{X}}, \phi_r, \alpha, \alpha_y) \tag{59}$$

$$\phi_y = \phi_y(\bar{X}, \dot{\bar{X}}, \phi_p, \alpha) \tag{60}$$

$$\ddot{\bar{X}} = \ddot{\bar{X}}(\bar{X}, \dot{\bar{X}}, \bar{\phi}, \alpha, \alpha_y) \tag{61}$$

$$H = H(\bar{X}, \dot{\bar{X}}, \alpha, \alpha_y, \bar{\phi}, \bar{\lambda}) \tag{62}$$

$$\dot{\bar{\lambda}} = \dot{\bar{\lambda}}(\bar{X}, \dot{\bar{X}}, \bar{\phi}, \bar{\lambda}, \alpha, \alpha_y) \tag{63}$$

$$\frac{\partial H}{\partial \alpha_y} = \frac{\partial H}{\partial \alpha_y}(\bar{X}, \dot{\bar{X}}, \alpha, \alpha_y, \bar{\phi}, \bar{\lambda}) = 0 \tag{64}$$

Starting values

ϕ_{ro}

m

GM

R_o

A

λ_{70}

$[\mu]$

$\lambda_{80} = 1$

$$\bar{X}_o = \begin{bmatrix} X_o \\ Y_o \\ Z_o \end{bmatrix} \qquad \bar{Z}_{ro} = \begin{bmatrix} 0 \\ 0 \\ Z_{ro} \end{bmatrix}$$

$$\dot{\bar{X}}_o = \begin{bmatrix} \dot{X}_o \\ \dot{Y}_o \\ \dot{Z}_o \end{bmatrix} \qquad \bar{\lambda}_{IO} = \begin{bmatrix} \lambda_{10} \\ \lambda_{20} \\ \lambda_{30} \end{bmatrix}$$

$$\bar{x}_{cp} = \begin{bmatrix} x_{cp} \\ y_{cp} \\ z_{cp} \end{bmatrix} \qquad \bar{\lambda}_{IIO} = \begin{bmatrix} \lambda_{40} \\ \lambda_{50} \\ \lambda_{60} \end{bmatrix}$$

Atmospheric tables for ρ as a function of altitude.

Atmospheric tables for $\bar{W}$ as a function of position.

Aerodynamic tables for $f(\alpha)$ as a function of α.

Preload Computation I

1. Choose $\alpha_{y1} = -\pi$

2. Compute the following, in order, using the positive sign in equation (58), the α_{y1} from step 1, and starting values.

α	from equation	(56)
F_r	from equation	(57)
$\dot{\phi}_r$	from equation	(58)
ϕ_p	from equation	(59)
ϕ_y	from equation	(60)
$\ddot{\bar{X}}$	from equation	(61)
H	from equation	(62)
$\frac{\partial H}{\partial \alpha_y}$	from equation	(64)

3. Choose $\alpha_{y2} = \alpha_{y1} + 5^{\circ}$ and repeat step 2 using α_{y2}; $\alpha_{y3} = \alpha_{y2} + 5^{\circ}$ and repeat step 2 using α_{y3}; etc., up to $\alpha_y = \pi$.

4. Repeat steps 1 - 3 using the negative sign in equation (58) rather than the positive sign.

The print out from preload I should be tabulated as follows:

EQN. 58+	α_y	H	$\frac{\partial H}{\partial \alpha_y}$	EQN. 58-	α_y	H	$\frac{\partial H}{\partial \alpha_y}$

Plots of H versus α_y and $\frac{\partial H}{\partial \alpha_y}$ versus α_y should give some insight as to whether more than one solution exists to this problem.

Preload Computation II

5. Using starting values and the positive sign in equation (58), iterate equation (64) for α_y. The results of preload I will aid in choosing the starting point for the iteration.

6. Compute $\frac{\partial^2 H}{\partial \alpha_y^2}$, equation (55), using starting values and the α_y from step 5.

7. Check for $\frac{\partial^2 H}{\partial \alpha_y^2} > 0$. If the inequality holds, then a minimum exists. Proceed to step 13 using the positive sign in equation (58) for all further calculations. If the inequality does not hold, proceed to step 8.

8. Check for $\frac{\partial^2 H}{\partial \alpha_y^2} \leq 0$. If this inequality holds, then go back and take the negative sign in equation (58).

9. Using starting values and the negative sign in equation (58), iterate equation (64) for α_y.

10. Compute $\frac{\partial^2 H}{\partial \alpha_y^2}$, equation (55), using starting values and the α_y from step 9.

11. Check to assure that $\frac{\partial^2 H}{\partial \alpha_y^2} > 0$.

12. Proceed to step 13 using the negative sign in equation (58) for all further calculations.

"N" Line Computation

13. Using starting values and the correct sign in equation (58), as chosen by preload II, iterate equation (64) for α_y. Use preload I to obtain the approximate iteration range.

14. Use the α_y from step 13 along with starting values to compute the following:

α	from equation	(56)
F_r	from equation	(57)
$\dot{\phi}_r$	from equation	(58)
ϕ_p	from equation	(59)
ϕ_y	from equation	(60)
$\ddot{\bar{X}}$	from equation	(61)
H	from equation	(62)
$\dot{\bar{\lambda}}_I$	from equation	(63)
$\dot{\bar{\lambda}}_{II}$	from equation	(63)
$\dot{\lambda}_7$	from equation	(63)

15. Use some numerical integration technique to integrate the following:

$\ddot{\bar{X}}$ for $\dot{\bar{X}}$ for $\bar{X}$

$\dot{\phi}_r$ for ϕ_r

$\dot{\bar{\lambda}}_I$ for $\bar{\lambda}_I$

$\dot{\bar{\lambda}}_{II}$ for $\bar{\lambda}_{II}$

$\dot{\lambda}_7$ for λ_7

16. Use integrated values from step 15 as starting values for the n + 1 line.

VIII. CONCLUSION

The problem analyzed in this paper has application in the fields of space flight and guidance. In space missions such as the earth-moon transit, the return to the earth's surface presents many problems. One of these problems is how to re-enter the earth's atmosphere with a wingless vehicle and make a point landing on the surface, at the same time minimizing the factors which cause strain on the human crew.

In this paper, the elimination of the oscillations due to first and second time derivatives of pitch and yaw and the second time derivative of roll is significant since, it is believed, these cause unnecessary strain on the vehicle's crew. This, in effect, replaces the dynamical motion of the altitude loop by its instantaneous steady state solution. The choice of α_y as the control variable seems physically realistic since this angle lies in a plane perpendicular to the roll axis and any change in this angle will be a roll of the vehicle about this axis. Such a control should allow maneuverability in three space.

In order to generate trajectories numerically, the initial auxiliary variables must be known. In this paper no attempt has been made to find these initial variables. It is assumed that they are known. If all initial values are assumed known, then trajectories generated numerically will satisfy the constraint and the characteristic equations. Satisfaction of the characteristic equations is a necessary but not sufficient condition for the existence of an optimum. A further necessary condition for the existence of a minimum is easily obtained from the Maximum

Principle, i.e., the condition that

$$\frac{\partial^2 H}{\partial \alpha_y^2} > 0$$

is necessary for a minimum of the integral D.

BIBLIOGRAPHY

Bliss, G. A. *Lectures on the Calculus of Variations*. Chicago: The University of Chicago Press, 1946.

Goldstein, Herbert. *Classical Mechanics*. Reading, Massachusetts: Addison-Wesley Publishing Company, Inc., 1959.

Kopp, Richard E. *Pontryagin Maximum Principle*, Chapter 7 of Optimization Techniques. Edited by George Leitmann. Berkeley, California: Academic Press, 1961.

Miner, W. E. Methods for Trajectory Computation, NASA-Marshall Space Flight Center, Internal Note, May 10, 1961.

Pontryagin, L. S., et al. *The Mathematical Theory of Optimal Processes*. New York: Interscience Publishers, 1962.

Progress Report No. 4 on Studies in the Fields of Space Flight and Guidance Theory, MTP-AERO-63-65. NASA-Marshall Space Flight Center, September 19, 1963.

APPENDICES

APPENDIX I

DISCUSSION ON WHY $[C]$ IS A NON-ZERO MATRIX

The matrix

$$[C] = \left\{ [A_\omega]^T \, [\mu] \, [A_\omega] \right\}^{-1}$$
$$= [A_\omega]^{-1} [\mu]^{-1} \left\{ [A_\omega]^T \right\}^{-1}$$

exists since it can be shown that the components of its product exist. The inverse of $[C]$ is developed below to show that it exists and is non-zero.

$$[C]^{-1} = \left[\begin{array}{l}
\mathrm{SRCY}(I_{xx}\mathrm{SRCY} - I_{xy}\ \mathrm{SY} - I_{xz}\ \mathrm{CRCY}) + \mathrm{SY}(-I_{xy}\ \mathrm{SRCY} + I_{yy}\ \mathrm{SY} \\
- I_{yz}\ \mathrm{CRCY}) - (I_{xy}\ \mathrm{SRCY} + I_{yy}\ \mathrm{SY} - I_{yz}\ \mathrm{CRCY}) \\
\mathrm{CR}(I_{xx}\ \mathrm{SRCY} - I_{xy}\ \mathrm{SY} - I_{xz}\ \mathrm{CRCY}) - \mathrm{SR}(-I_{xz}\ \mathrm{SRCY} - I_{yz}\ \mathrm{SY}
\end{array}\right.$$

$$\begin{array}{l:l:}
+ \mathrm{CRCY}(-I_{xz}\ \mathrm{SRCY} - I_{yz}\ \mathrm{SY} + I_{zz}\ \mathrm{CRCY} & \mathrm{SR}\ I_{xy} - \mathrm{SY}\ I_{yy} + \mathrm{CRCY}\ I_{yz} \\
+ I_{zz}\ \mathrm{CRCY}) & \\
 & I_{yy} \\
 & \mathrm{CR}\ I_{xy} - \mathrm{SR}\ I_{yz}
\end{array}$$

$$\left.\begin{array}{:l}
\mathrm{SRCY}(I_{xx}\ \mathrm{CR} + I_{xz}\ \mathrm{SR}) + \mathrm{SY}(-I_{xy}\ \mathrm{CR} + I_{yz}\ \mathrm{SR}) - \mathrm{CRCY}(I_{xz}\ \mathrm{CR} + I_{zz}\ \mathrm{SR}) \\
\\
I_{xy}\ \mathrm{CR} - I_{yz}\ \mathrm{SR} \\
\\
\mathrm{CR}(I_{xx}\ \mathrm{CR} + I_{xz}\ \mathrm{SR}) + \mathrm{SR}\ (I_{xz}\ \mathrm{CR} + I_{zz}\ \mathrm{SR})
\end{array}\right]$$

Thus from the definition

$$[C]\,[C]^{-1} = [I]$$

it follows that $[C]$ is a non-zero matrix.

APPENDIX II

UNIQUENESS OF SOLUTION FOR ϕ_p

The expressions for sine ϕ_p and cosine ϕ_p were found by using the quadratic formula. The radical thus carries the sign ± . In order for $S^2P + C^2P = 1$ the plus sign must be chosen with the sine radical and the minus sign must be chosen with the cosine radical or vice versa. Either combination will give a solution for ϕ_p . The unique solution is chosen from these two by considering the way in which the coordinate systems were defined. Consider

$$SP = \frac{J\, V_{RY} \pm \sqrt{J^2 V_{RY}^2 - (V_{RX}^2 + V_{RY}^2)(J^2 - V_{RX}^2)}}{(V_{RX}^2 + V_{RY}^2)}$$

where $J = CR\, V_{rmx} - SR\, V_{rmz}$.

Let $\phi_r = \alpha = 0$. This implies $J = 0$.

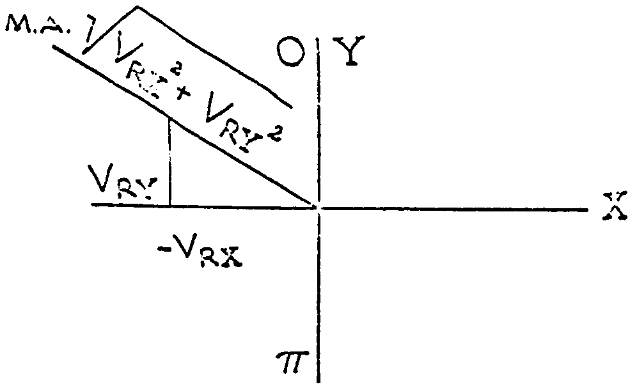
Then,

$$SP = \pm \frac{V_{RX}}{\sqrt{V_{RX}^2 + V_{RY}^2}}$$

and

$$\tan \phi_p = \frac{-V_{RX}}{V_{RY}}$$

Now restrict ϕ_p, $-\pi \le \phi_p \le \pi$



The correct signs are thus,

$$SP = \frac{-V_{RX}}{\sqrt{V_{RX}^2 + V_{RY}^2}} \quad \text{and} \quad CP = \frac{+V_{RY}}{\sqrt{V_{PX}^2 + V_{RY}^2}} \quad .$$

APPENDIX III

UNIQUENESS OF SOLUTION FOR ϕ_y

The expressions for sine ϕ_y and cosine ϕ_y were found by using the quadratic formula. The radical thus carries the sign ± . In order for $S^2Y + C^2Y = 1$, the plus sign must be chosen with the sine radical and the minus sign must be chosen with the cosine radical or vice versa. Either combination will give a solution for ϕ_y. The unique solution is chosen from these two by considering the way in which the coordinate systems were defined.

Consider

$$SY = \frac{V_{rmy} V_{RZ} \pm \sqrt{V_{rmy}^2 V_{RZ}^2 - (V_{RZ}^2 + K^2)(V_{rmy}^2 - K^2)}}{(V_{RZ}^2 + K^2)}$$

where $K = CP\, V_{RY} - SP\, V_{RX}$

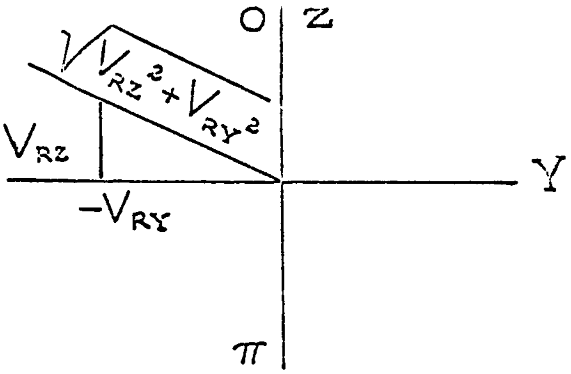
Let $\phi_p = 0$. This implies $K = V_{RY}$

Let $\alpha = 90^\circ$. This implies $V_{rmy} = 0$.

Then, $SY = \dfrac{\pm V_{RY}}{\sqrt{V_{RZ}^2 + V_{RY}^2}}$

and $\tan \phi_y = \dfrac{-V_{RY}}{V_{RZ}}$.

Now restrict ϕ_y, $-\pi \le \phi_y \le \pi$.



The correct signs are thus,

$$SY = \frac{-V_{RY}}{\sqrt{V_{RZ}^2 + V_{RY}^2}} \quad \text{and} \quad CY = \frac{+V_{RZ}}{\sqrt{V_{RZ}^2 + V_{RY}^2}} .$$

AN OPTIMAL GUIDANCE APPROXIMATION THEORY

Henry J. Kelley
Analytical Mechanics Associates, Inc.

January 1964

AN OPTIMAL GUIDANCE APPROXIMATION THEORY*

Henry J. Kelley**
Analytical Mechanics Associates, Inc.

SUMMARY

20953

A

Synthesis of optimal guidance approximations is undertaken by means of a perturbation theory approach. A simple example is treated analytically and an approximation for the optimal control including linear and quadratic feedback terms in the state deviations from an optimal reference trajectory obtained. Author

INTRODUCTION

The problem of guidance in the neighborhood of an optimized nominal trajectory has previously been studied from slightly differing viewpoints by Kelley (Ref. 1) and Breakwell and Bryson (Ref. 2) who have developed a procedure for synthesizing linear feedback guidance approximations optimal in the same sense as the nominal trajectory. The present paper deals with synthesis of higher order approximations by means of perturbation theory applied to the Euler-Lagrange equations, and presents a transparently simple illustrative example in which quadratic feedback terms can be calculated analytically.

* This research was performed under Contract NAS 8-5314 with the Aerodynamics and Astrodynamics Division of NASA Marshall Space Flight Center, Huntsville, Alabama.

** Vice-President

PROPERTIES OF THE NOMINAL TRAJECTORY

The nominal trajectory is assumed to satisfy the equations of state

$$\dot{x}_i = g_i(x_1, - -, x_n, y_1, - -, y_\ell, t) \qquad i = 1, - -, n \tag{1}$$

and the Euler-Lagrange equations

$$\dot{\lambda}_i = - \frac{\partial H}{\partial x_i} , \qquad i = 1, - -, n \tag{2}$$

$$\frac{\partial H}{\partial y_k} = 0 , \qquad k = 1, - -, \ell \tag{3}$$

subject to boundary conditions at the initial and terminal points consisting of appropriate specified conditions and transversality conditions. The function

$$H \equiv \sum_{i=1}^{n} \lambda_i g_i \tag{4}$$

is the usual Hamiltonian and the λ_i are Lagrange multipliers.

It is assumed that the reference solution of (1), (2) and (3) which represents the optimized nominal trajectory provides a minimum of a function $P(x_{1_f}, - -, x_{n_f}, t_f)$ of the terminal values and is a normal nonsingular extremal without corners which satisfies the strengthened forms of the Weierstrass and generalized Jacobi conditions. While some of these assumptions are introduced merely for convenience, and can be relaxed, others, such as the requirements of nonsingularity and nonconjugate end-points, are essential to the development following.

PERTURBATION THEORY

For simplicity of exposition, the system of differential equations (2) and (3) will be written in the form

$$\dot{z}_p = f_p(z_1, --, z_{2n}, t) \qquad (5)$$

$$p = 1, --, 2n$$

in which $z_1, --, z_n$ are identified as $x_1, --, x_n$ and $z_{n+1}, --, z_{2n}$ as $\lambda_1, --, \lambda_n$, and the control variables $y_1, --, y_\ell$ have been presumed eliminated by use of the Weierstrass condition. This will always be possible within the framework of our assumptions since the strengthened Weierstrass condition implies an unique minimum of $H(y_1, --, y_\ell)$. In practical applications it may be preferable to retain the control variables and eqs. (3), but this will necessitate no essential change in the arguments to follow.

It is desired to develop an approximation to the family of solutions of the system (5) in the vicinity of the reference solution corresponding to the nominal trajectory. The parameters of the family will be the deviations of the initial state variables from their reference initial values

$$\epsilon_q = z_q(t_o) - \bar{z}_q(t_o) \quad , \qquad q = 1, --, n \qquad (6)$$

and the family will be represented in terms of a Taylor series expansion

$$z_p = \bar{z}_p + \sum_{r=1}^{n} \frac{\partial z_p}{\partial \epsilon_r} \epsilon_r + \frac{1}{2} \sum_{r,s=1}^{n} \frac{\partial^2 z_p}{\partial \epsilon_r \partial \epsilon_s} \epsilon_r \epsilon_s + --- \qquad (7)$$

$$p = 1, --, 2n$$

Here, and in (6), the superscribed bar signifies the reference solution of (5).

If (7) is introduced into the system of differential equations (5) and the right members of (5) expanded in the ϵ_q, the resulting expression

$$\frac{d}{dt}\left[\bar{z}_p + \sum_{r=1}^{n} \frac{\partial z_p}{\partial \epsilon_r}\epsilon_r + \frac{1}{2}\sum_{r,\,s=1}^{n} \frac{\partial^2 z_p}{\partial \epsilon_r \partial \epsilon_s}\epsilon_r\epsilon_s + ---\right] =$$

$$f_p(\bar{z}) + \sum_{q=1}^{2n} \frac{\partial f_p}{\partial z_q}\left[\sum_{r=1}^{n} \frac{\partial z_q}{\partial \epsilon_r}\epsilon_r\right] + \frac{1}{2}\sum_{q=1}^{2n} \frac{\partial f_p}{\partial z_q}\left[\sum_{r,\,s=1}^{n} \frac{\partial^2 z_q}{\partial \epsilon_r \partial \epsilon_s}\epsilon_r\epsilon_s\right]$$

$$+ \frac{1}{2}\sum_{q,\,u=1}^{2n} \frac{\partial^2 f_p}{\partial z_q \partial z_u}\left[\sum_{r=1}^{n} \frac{\partial z_q}{\partial \epsilon_r}\epsilon_r\right]\left[\sum_{r=1}^{n} \frac{\partial z_u}{\partial \epsilon_r}\epsilon_r\right] + --- \qquad (8)$$

$$p = 1, --, 2n$$

may be regarded as an identity in the parameters ϵ_q. This leads to a system of differential equations governing the partial derivatives which are the coefficients in (7):

$$\frac{d}{dt}\frac{\partial z_p}{\partial \epsilon_r} = \sum_{q=1}^{n} \frac{\partial f_p}{\partial z_q}\frac{\partial z_q}{\partial \epsilon_r}, \qquad \begin{array}{l} p = 1, --, 2n \\ r = 1, --, n \end{array} \qquad (9)$$

$$\frac{d}{dt}\frac{\partial^2 z_p}{\partial \epsilon_r \partial \epsilon_s} = \sum_{q=1}^{2n} \frac{\partial f_p}{\partial z_q}\frac{\partial^2 z_q}{\partial \epsilon_r \partial \epsilon_s} + \sum_{q,\,u=1}^{2n} \frac{\partial^2 f_p}{\partial z_q \partial z_u}\left[\frac{\partial z_q}{\partial \epsilon_r}\frac{\partial z_u}{\partial \epsilon_s} + \frac{\partial z_q}{\partial \epsilon_s}\frac{\partial z_u}{\partial \epsilon_r}\right] \qquad (10A)$$

$$\begin{array}{l} p = 1, --, 2n \\ r = 1, --, n \neq s \\ s = 1, --, n \end{array}$$

$$\frac{d}{dt}\frac{\partial^2 z_p}{\partial \epsilon_r \partial \epsilon_r} = \sum_{q=1}^{2n} \frac{\partial f_p}{\partial z_q}\frac{\partial^2 z_q}{\partial \epsilon_r \partial \epsilon_r} + \sum_{q,\,u=1}^{2n} \frac{\partial^2 f_p}{\partial z_q \partial z_u}\frac{\partial z_q}{\partial \epsilon_r}\frac{\partial z_u}{\partial \epsilon_r} \qquad (10B)$$

$$\begin{array}{l} p = 1, --, 2n \\ r = 1, --, n \end{array}$$

The process may be carried out to obtain differential equations for the partial derivatives of any order occurring in (7) provided that the functions f_p are smooth enough to permit the required differentiations.

If numerical procedures are intended, it will usually be desirable to work with the partial derivative coefficients and the differential equations (9) and (10), while, on the other hand, for analytical treatments it will often be convenient to introduce the linear combinations

$$\delta z_p \equiv \sum_{r=1}^{n} \frac{\partial z_p}{\partial \epsilon_r} \epsilon_r \quad , \qquad p = 1, --, 2n \tag{11}$$

$$\delta^2 z_p \equiv \frac{1}{2} \sum_{r,s=1}^{n} \frac{\partial^2 z_p}{\partial \epsilon_r \partial \epsilon_s} \epsilon_r \epsilon_s \quad , \qquad p = 1, --, 2n \tag{12}$$

which satisfy the systems

$$\delta \dot{z}_p = \sum_{q=1}^{2n} \frac{\partial f_p}{\partial z_q} \delta z_q \quad , \qquad p = 1, --, 2n \tag{13}$$

$$\delta^2 \dot{z}_p = \sum_{q=1}^{2n} \frac{\partial f_p}{\partial z_q} \delta^2 z_q + \frac{1}{2} \sum_{q,u=1}^{2n} \frac{\partial^2 f_p}{\partial z_q \partial z_u} \delta z_q \delta z_u \tag{14}$$

$$p = 1, --, 2n$$

The boundary conditions applying at the terminal point of the trajectory are of the general form

$$\Psi_q(z_1, --, z_{2n}, t)_{t_f} = 0 \quad , \qquad q = 1, --, n+1 \tag{15}$$

and these may similarly be expanded as

$$\begin{aligned}\Psi_q(\bar{z}_{1_f}, --, \bar{z}_{2n_f}, \bar{t}_f) &+ \Big\{\sum_{s=1}^{2n} \frac{\partial \Psi_q}{\partial z_s}\Big[\delta z_s + f_s \delta t + \delta^2 z_s + f_s \delta^2 t + \sum_{u=1}^{2n} \frac{\partial f_s}{\partial z_u}(\delta z_u + f_u \frac{\delta t}{2})\delta t \\ &+ \frac{\partial f_s}{\partial t}\,\frac{\delta t^2}{2}\Big] + \frac{\partial \Psi_q}{\partial t}(\delta t + \delta^2 t) + \frac{1}{2}\sum_{s,u=1}^{2n} \frac{\partial^2 \Psi_q}{\partial z_s \partial z_u}(\delta z_s + f_s \delta t)(\delta z_u + f_u \delta t) \\ &+ \sum_{s=1}^{2n} \frac{\partial^2 \Psi_q}{\partial z_s \partial t}(\delta z_s + f_s \delta t)\delta t + \frac{1}{2}\,\frac{\partial^2 \Psi_q}{\partial t^2}\,\delta t^2 + ---\Big\}_{\bar{t}_f} = 0 \qquad (16)\end{aligned}$$

$$q = 1, --, n+1$$

The symbols δt and $\delta^2 t$ appearing here are the first and second variations in the terminal time, defined by

$$\delta t_f = \sum_{r=1}^{n} \frac{\partial t_f}{\partial \epsilon_r}\epsilon_r \qquad (17)$$

$$\delta^2 t_f = \frac{1}{2}\sum_{r,s=1}^{n} \frac{\partial^2 t_f}{\partial \epsilon_r \partial \epsilon_s}\epsilon_r \epsilon_s \qquad (18)$$

and the various partial derivatives appearing are evaluated at the terminal point of the nominal trajectory at the nominal terminal time $\bar{t}_f$.

At the fixed initial time t_0, the initial conditions are given by (6), which, with the introduction of (7), may be regarded as identities in the parameters ϵ_r, as also may the terminal conditions (16). Thus boundary conditions may be derived for the system (9) and (10) in the partial derivatives, or, alternatively, for the system (13) and (14) in the variations.

COMPUTATIONAL CONSIDERATIONS

If only first order terms of the expansion are sought, as in Refs. 1 and 2, the linearity of the system (9) or (13) may be exploited by the introduction of the corresponding adjoint system, by means of which the expansion coefficients may be calculated economically over a range of "initial" times extending from the initial to the terminal time of the nominal trajectory. No such economy measure is available, however, in the computation of second and higher order terms, for while the system (10) or (14) is linear, the nonhomogeneous terms are quadratic functions of the first order solution, and hence the systems (9) and (10), or (13) and (14), viewed as a simultaneous system, are nonlinear, and the adjoint device is inapplicable.

ANALYTICAL TREATMENT OF AN EXAMPLE: ZERMELO'S PROBLEM

The simple problem for which the linear feedback terms were calculated in Ref. 1 affords the possibility of obtaining the quadratic feedback terms analytically as well. A particle moves with constant speed V relative to a medium which itself is in motion with velocity components u and v, presumed constant. The equations of state are

$$\dot{z} = V \sin\gamma + u \tag{19}$$

$$\dot{x} = V \cos\gamma + v , \tag{20}$$

The steering angle γ is the control variable of the problem, and the minimum time path from a specified initial point to a fixed destination point (z^*, x^*) is sought. The extremals are straight lines and collision guidance is optimal.

The Euler-Lagrange equations are

$$\dot{\lambda}_1 = 0 \tag{21}$$

$$\dot{\lambda}_2 = 0 \tag{22}$$

$$\lambda_1 \cos\gamma - \lambda_2 \sin\gamma = 0 \tag{23}$$

and the optimal steering angle γ is determined by

$$\sin\gamma = \frac{-\lambda_1}{(\lambda_1^2+\lambda_2^2)^{1/2}} \quad , \qquad \cos\gamma = \frac{-\lambda_2}{(\lambda_1^2+\lambda_2^2)^{1/2}} \tag{24}$$

The transversality condition

$$H(t_f) = -1 \tag{25}$$

applies at the terminal point.

The numerical data for the example and for the path chosen as a nominal trajectory are

$$\begin{array}{lll} \bar{z}_o = 0 & \bar{z}_f = z^* = 1 & \bar{t}_o = 0 \\ \bar{x}_o = 0 & \bar{x}_f = x^* = 2 & \bar{t}_f = 2 \\ V = 1 & u = 1/2 & v = 0 \\ \bar{\lambda}_1 = 0 & \bar{\lambda}_2 = -\frac{1}{V} = -1 & \\ \sin\bar{\gamma} = 0 & \cos\bar{\gamma} = 1 & \end{array} \tag{26}$$

The equations for the first order guidance solution are

$$\delta\dot{z} = V \cos\bar{\gamma}\, \delta\gamma \tag{27}$$

$$\delta\dot{x} = -V \sin\bar{\gamma}\, \delta\gamma \tag{28}$$

$$\delta\dot{\lambda}_1 = 0 \tag{29}$$

$$\delta\dot{\lambda}_2 = 0 \tag{30}$$

$$\delta\lambda_1 \cos\bar{\gamma} - \delta\lambda_2 \sin\bar{\gamma} - (\lambda_1 \sin\bar{\gamma} + \lambda_2 \cos\bar{\gamma})\delta\gamma = 0 \tag{31}$$

These are subject to boundary conditions

$$\delta z(t_o) = z(t_o) - \bar{z}(t_o) \tag{32}$$

$$\delta x(t_o) = x(t_o) - \bar{x}(t_o) \tag{33}$$

at the initial point and

$$\delta z(\bar{t}_f) + (V \sin\bar{\gamma}_f + u)\delta t_f = 0 \tag{34}$$

$$\delta x(\bar{t}_f) + (V \cos\bar{\gamma}_f + v)\delta t_f = 0 \tag{35}$$

$$\delta\lambda_1(\bar{t}_f)(V \sin\bar{\gamma}_f + u) + \delta\lambda_2(\bar{t}_f)(V \cos\bar{\gamma}_f + v) = 0 \tag{36}$$

at the terminal point. Equations (34), (35) and (36) are obtained from the vanishing of the first order terms of the general expression (16).

The equations for the second order guidance solution are

$$\delta^2 \dot{z} = V \cos\bar{\gamma}\, \delta^2\gamma - V \sin\bar{\gamma}\, \frac{\delta\gamma^2}{2} \tag{37}$$

$$\delta^2 \dot{x} = - V \sin\bar{\gamma}\, \delta^2\gamma - V \cos\bar{\gamma}\, \frac{\delta\gamma^2}{2} \tag{38}$$

$$\delta^2 \dot{\lambda}_1 = 0 \tag{39}$$

$$\delta^2 \dot{\lambda}_2 = 0 \tag{40}$$

$$\delta^2\lambda_1 \cos\bar{\gamma} - \delta^2\lambda_2 \sin\bar{\gamma} - (\lambda_1 \sin\bar{\gamma} + \lambda_2 \cos\bar{\gamma})\delta^2\gamma$$

$$+ (-\lambda_1 \cos\bar{\gamma} + \lambda_2 \sin\bar{\gamma})\frac{\delta\gamma^2}{2} = 0 \tag{41}$$

These are subject to boundary conditions

$$\delta^2 z(t_o) = 0 \tag{42}$$

$$\delta^2 x(t_o) = 0 \tag{43}$$

at the initial point and

$$\delta^2 z(\bar{t}_f) + V \cos\bar{\gamma}_f\, \delta\gamma(\bar{t}_f)\delta t_f + (V \sin\bar{\gamma}_f + u)\delta^2 t_f = 0 \tag{44}$$

$$\delta^2 x(\bar{t}_f) - V \sin\bar{\gamma}_f\, \delta\gamma(\bar{t}_f)\delta t_f + (V \cos\bar{\gamma}_f + v)\delta^2 t_f = 0 \tag{45}$$

$$\delta^2\lambda_1(\bar{t}_f)(V \sin\bar{\gamma}_f + u) + \delta^2\lambda_2(\bar{t}_f)(V \cos\bar{\gamma}_f + v) - \frac{V}{2\sqrt{\bar{\lambda}_{1_f}^2 + \bar{\lambda}_{2_f}^2}}\,[\delta\lambda_1(\bar{t}_f)\cos\bar{\gamma}_f - \delta\lambda_2(\bar{t}_f)\sin\bar{\gamma}_f]^2 = 0 \tag{46}$$

at the terminal point. Equations (44), (45) and (46) are obtained from the vanishing of the second order terms of the general expression (16).

The first order guidance solution is that given in Ref. 1:

$$\delta\gamma = -\frac{\delta z_o}{V(\bar{t}_f - t_o)} + \frac{u\,\delta x_o}{V^2(\bar{t}_f - t_o)} \tag{47}$$

$$\delta t_f = -\frac{\delta x_o}{V} \tag{48}$$

in which the simplifications $\sin\bar{\gamma} = 0$, $\cos\bar{\gamma} = 1$, $v = 0$ of the specific examples have been introduced. The corresponding second order results are

$$\delta^2\gamma = \frac{\delta x_o\,\delta\gamma}{V(\bar{t}_f - t_o)} - \frac{u}{2V}\,\delta\gamma^2 \tag{49}$$

$$\delta^2 t_f = \frac{\delta\gamma^2}{2} (\bar{t}_f - t_o) \tag{50}$$

and the guidance law incorporating both first and second order terms is

$$\delta\gamma + \delta^2\gamma = - \frac{\delta z}{V(\bar{t}_f - t)} + \frac{u\delta x}{V^2(\bar{t}_f - t)} - \frac{u}{2V}\left[\frac{-\delta z}{V(\bar{t}_f - t)} + \frac{u\delta x}{V^2(\bar{t}_f - t)}\right]^2$$

$$+ \frac{\delta x}{V(\bar{t}_f - t)}\left[\frac{-\delta z}{V(\bar{t}_f - t)} + \frac{u\,\delta x}{V^2(\bar{t}_f - t)}\right] \tag{51}$$

In this expression, t_o has been replaced by instantaneous time t, as appropriate for continuous closed-loop system operation.

CONCLUDING REMARKS

A computer simulator study is in progress to determine the effects of the second order feedback terms of the example upon system performance and guidance accuracy.

REFERENCES

1. Kelley, H. J.; "Guidance Theory and Extremal Fields," IRE National Aerospace Electronics Conference, May 14-16, 1962; also IRE Transactions on Automatic Control, October 1962.

2. Breakwell, J. V. and Bryson, A. E.; "Neighboring-Optimum Terminal Control for Multivariable Nonlinear Systems," S. I. A. M. Symposium on Multivariable System Theory, Cambridge, Mass., Nov. 1-3, 1962; also Breakwell, J. V., Speyer, J. L. and Bryson, A. E.; "Optimization and Control of Nonlinear Systems Using the Second Variation," S. I. A. M. Journal on Control, Vol. 1, No. 2, 1963.

20954

CONSTANTS OF THE MOTION FOR OPTIMUM THRUST TRAJECTORIES IN A CENTRAL FORCE FIELD

Samuel Pines

October 1, 1963

Analytical Mechanics Associates, Inc.
941 Front Street
Uniondale, L. I. N. Y.

ACKNOWLEDGEMENT

This study was carried out by Analytical Mechanics Associates, Inc. under contract NAS 8-5314 for the Future Projects Branch, Aeroballistics Division.

SUMMARY

20954

A

This paper derives four constants of the motion for optimal thrust trajectories in a central force field. Two additional constants of the motion are derived which hold for singular thrusting arcs as well as impulsive thrusts.

The paper applies the constants of the motion for the impulsive thrust case, to obtain a set of initial conditions for the classical adjoint variables to be used ac a good approximation for a solution of the finite thrust arc by the indirect method.

Author

INTRODUCTION

The constants of the motion of a system of differential equations play an important role in characterizing the solutions. This paper develops an application of the constants of the motion to the indirect methods for obtaining solutions of optimal thrust trajectories by iterative procedures.

The optimal trajectories for a thrusting vehicle in a central force field have been under study for some time by Lawden[1], Leitmann[2], Melbourne[3], Breakwell[4], and others. Four constants of the motion for this problem are well known. This paper derives two additional constants of the motion which hold for singular thrusting arcs and impulsive thrusts. The paper also derives the four known constants of the motion. The paper applies the constants of the motion for the impulsive thrust case to obtain a set of initial conditions for the classical adjoint variables to be used as a good approximation for the solution of the finite thrust arc by the indirect method.

As is well known, the indirect methods for obtaining solutions of the optimal thrust trajectories by iterative procedures suffer from an extreme sensitivity of the solution to small changes in the initial conditions of the adjoint variables. In effect, the success of the gradient techniques, employed by Kelley[5] and Bryson[6], is largely due to their ability to control the incremental step size for small changes in the thrusting logic.

Once a good approximation to the optimum control thrust logic has been obtained, the gradient techniques prove too slow for convergence and resort is made to the classical indirect methods for the last few iterations. If a good approximation to the initial conditions of the adjoint variables were available, the indirect methods would be in more general use.

A good approximation to the initial conditions of the adjoint variables, for a given problem may be obtained through a study of the limiting impulsive solution to the same problem. Let us assume that a solution to an optimal thrust trajectory exists and that it is known. Then, if one could improve the efficiency of the engine (thrust/weight ratio), a shorter burning arc could be obtained for an improved optimal trajectory. In the limit one would obtain the impulsive thrust solution of the given problem which would indeed require a perfect engine. Thus, we can look at the impulsive solution as a limiting point in a simply connected region in the space of the initial conditions of the adjoint variables. Intuitively, one might expect that an iterative procedure could be developed which would start with the known impulsive solution and converge to the required finite thrust solution.

This report applies the constants of the motion for an optimal impulsive trajectory to obtain approximate values of the adjoint variables to be used for an indirect method solution of the optimal trajectory with finite thrust. The specific example illustrated in this report is the minimum fuel required for bounded thrust between fixed initial and final position and velocity states.

I. The Equations of Motion

The equations of motion of a thrusting vehicle in a central force field are given by

$$\begin{aligned} \ddot{R} &= -\mu \frac{R}{r^3} + \frac{k}{m} T \\ \dot{m} &= -\frac{k}{c} , \end{aligned} \tag{1}$$

where $|T| = 1$, and $c = \text{constant.}$

The necessary conditions for minimizing the fuel consumption with time open, or minimum time for fixed fuel, are given by

$$T = \frac{\lambda}{|\lambda|} , \tag{2}$$

and

$$\begin{aligned} k &= k_{max} && \text{if} && \left(|\lambda| - \frac{m\sigma}{c} > 0 \right) \\ k &= k_{min} && \text{if} && \left(|\lambda| - \frac{m\sigma}{c} < 0 \right) \\ k_{min} \leq k &\leq k_{max} && \text{if} && \left(|\lambda| \equiv \frac{m\sigma}{c} \right) . \end{aligned} \tag{3}$$

The adjoint variables are solutions of the Euler-Lagrange differential equations as follows:

$$\begin{aligned} \ddot{\lambda} &= -\mu \frac{\lambda}{r^3} + 3\mu \frac{\lambda \cdot R}{r^3} R \\ \dot{\sigma} &= \frac{k}{m^2} \lambda \cdot T . \end{aligned} \tag{4}$$

The final equations which yield the optimum trajectories (if any exist at all) are given by

$$\ddot{R} = -\mu \frac{R}{r^3} + \frac{k}{m} \frac{\lambda}{|\lambda|} , \qquad (5a)$$

$$\ddot{\lambda} = -\mu \frac{\lambda}{r^3} + 3\mu \frac{R \cdot \lambda}{r^5} R , \qquad (5b)$$

$$\dot{m} = -\frac{k}{c} , \qquad (5c)$$

$$\dot{\sigma} = \frac{k}{m^2} |\lambda| . \qquad (5d)$$

To obtain the proper solution, it is necessary to make some statement about the initial and final conditions of the state variables. For the purposes of this paper it will be sufficient to characterize all the solutions of the equations of motion through the constants of the motion. For this reason no further discussion of the initial, final, or transversality conditions will be carried out.

II. The Constants of the Motion

This section contains a derivation of four constants of the motion of equations (5a) - (5d). For the special case of the singular thrusting arc and the impulsive solutions, two more constants of the motion are given.

By forming the vector cross product of λ with equation (5a), the vector cross product of R with equation (5b), and adding, the following equation results

$$\lambda \times \ddot{R} + R \times \ddot{\lambda} = - \mu \frac{\lambda \times R}{r^3} - \mu \frac{R \times \lambda}{r^3} = 0 . \tag{6}$$

Thus, three constants of the motion are given by the vector equation

$$\frac{d}{dt} (\lambda \times \dot{R} + R \times \dot{\lambda}) = 0 \tag{7}$$

The equation may be written as a vector constant

$$\lambda \times \dot{R} + R \times \dot{\lambda} = A . \tag{8}$$

In order to obtain a more convenient form for the optimal thrust logic, equations (5c) and (5d) may be combined as follows:

$$\begin{aligned} \frac{d}{dt} (m\sigma) &= \dot{m}\sigma + m\dot{\sigma} \\ &= - k \frac{\sigma}{c} + k \frac{|\lambda|}{m} . \end{aligned} \tag{9}$$

Thus

$$\frac{d}{dt} (m\sigma) = \frac{k}{m} (|\lambda| - \frac{m\sigma}{c}) . \tag{10}$$

If the coefficient of k is positive we use k_{max}, if the coefficient is negative we use k_{min}. In either case k is a constant so long as its coefficient is not identically zero. On the other hand, if the coefficient of k is identically zero, then

$$\frac{d}{dt}(m\sigma) \equiv 0 \ . \tag{11}$$

This condition is satisfied along a singular arc so that

$$m\sigma = \text{constant} \ . \tag{12}$$

Since $|\lambda| - \frac{m\sigma}{c} \equiv 0$, it follows that $|\lambda|$ is a constant. Thus the optimum thrust logic may be stated simply as follows: either

$$\begin{aligned} &k = \text{constant} && \text{if} && |\lambda| - \frac{m\sigma}{c} \neq 0 \\ \text{or} \\ &|\lambda| = \text{constant} && \text{if} && |\lambda| - \frac{m\sigma}{c} \equiv 0 \end{aligned} \tag{13}$$

It is now possible to obtain the fourth constant of the motion. Form the dot product of equation (5a) with $\dot{\lambda}$, the dot product of equation (5b) with $\dot{R}$, and add. The result is

$$\begin{aligned} \dot{\lambda} \cdot \ddot{R} + \dot{R} \cdot \ddot{\lambda} &= \frac{d}{dt}(\dot{\lambda} \cdot \dot{R}) \\ &= -\mu \frac{d}{dt}\left(\frac{\lambda \cdot R}{r^3}\right) + \frac{k}{m}\frac{d}{dt}|\lambda| \ . \end{aligned} \tag{14}$$

Since

$$\begin{aligned} k \cdot \frac{d}{dt}\frac{|\lambda|}{m} &= \frac{k}{m}\frac{d}{dt}|\lambda| + k\,|\lambda|\frac{d}{dt}\left(\frac{1}{m}\right) \\ \text{and} \\ k\frac{d}{dt}\frac{\sigma}{c} &= k\,|\lambda|\frac{d}{dt}\left(\frac{1}{m}\right) , \end{aligned} \tag{15}$$

it follows that

$$k \frac{d}{dt} \left(\frac{|\lambda|}{m} - \frac{\sigma}{c} \right) = \frac{k}{m} \frac{d}{dt} |\lambda|$$

$$= \frac{d^2}{dt^2} m\sigma \,. \tag{16}$$

Thus a fourth constant of the motion is given by

$$\dot{\lambda} \cdot \dot{R} + \mu \frac{R \cdot \lambda}{r^3} - \frac{d}{dt} m\sigma = h \tag{17}$$

From equation (10), an altered form of this constant of the motion is

$$\dot{\lambda} \cdot \dot{R} + \mu \frac{R \cdot \lambda}{r^3} - k \left(\frac{|\lambda|}{m} - \frac{\sigma}{c} \right) = h \tag{17a}$$

This is the so-called Hamiltonian. In particular, for the singular arc

$$\frac{d}{dt} m\sigma = 0 \,,$$

and (18)

$$\dot{\lambda} \cdot \dot{R} + \mu \frac{R \cdot \lambda}{r^3} = h \,.$$

Another constant of the motion may be obtained for some restricted cases. Form the dot product of equation (5a) with λ, the dot product of equation (5b) with R and subtract. The result is

$$\lambda \cdot \ddot{R} - R \cdot \ddot{\lambda} = \frac{d}{dt} (\lambda \cdot \dot{R} - R \cdot \dot{\lambda})$$

$$= - 3\mu \frac{\lambda \cdot R}{r^3} + \frac{k}{m} |\lambda| \tag{19}$$

In addition , we have

$$\begin{aligned} \frac{d}{dt}\,(\dot{\lambda}\cdot R) &= \dot{\lambda}\cdot\dot{R} + \ddot{\lambda}\cdot R \\ &= \dot{\lambda}\cdot\dot{R} + 2\mu\,\frac{\lambda\cdot R}{r^3} \end{aligned} \qquad (20)$$

It is possible to eliminate $\dot{\lambda}\cdot\dot{R}$ between equation (18) and equation (20) as follows:

$$\dot{\lambda}\cdot\dot{R} = \frac{d}{dt}\,(\dot{\lambda}\cdot R) - 2\mu\,\frac{\lambda\cdot R}{r^3} = h - \mu\,\frac{\lambda\cdot R}{r^3} + \frac{d}{dt}\,(m\sigma) \qquad (21)$$

It is also possible to eliminate $\mu\,\frac{\lambda\cdot R}{r^3}$ between equation (19) and (21) as follows:

$$\frac{1}{3}\,\frac{d}{dt}\,(\lambda\cdot\dot{R} - R\cdot\dot{\lambda}) + \frac{|\lambda|c}{3}\,\frac{d}{dt}\,(\log m) = \frac{d}{dt}\,(m\sigma + ht - \dot{\lambda}\cdot R) \qquad (22)$$

If $|\lambda|$ is a constant (this is the case for the singular arc) we have as a fifth constant of the motion

$$\frac{1}{3}\,\lambda\cdot\dot{R} + \frac{2}{3}\,R\cdot\dot{\lambda} + \frac{|\lambda|c}{3}\,\log m - ht = b\ . \qquad (23)$$

Moreover, for the same restrictive case, another constant of the motion is given by

$$m\sigma = d. \qquad (24)$$

To obtain the form of these new constants of the motion for impulsive thrust, some care must be taken in approaching the limit forms.

It is necessary to distinguish between impulsive thrusts in the interior time domain between the initial and final conditions, and the impulsive thrusts at the boundaries.

a) Interior impulsive thrust.

From equation (16) we have

$$- c \frac{d}{dt} \ (\log m) \frac{d}{dt} \ |\lambda| = \frac{d^2}{dt^2} \ (m\sigma) \tag{16a}$$

Integrating over an interior impulse, we have

$$-c \ \{ \log (m^+) - \log (m^-) \} \ \frac{d}{dt} \ |\lambda| = (\frac{d}{dt} \ m\sigma)^+ - \frac{d}{dt}(m\sigma)^- \tag{25}$$

It is plain that during an interior coasting arc, the engine is off, $k = 0$, and $\frac{d}{dt}$ $(m\sigma) = 0$. Thus, it follows

$$-c \ \{ \log (m^+) - \log (m^-) \} \ \frac{d}{dt} \ |\lambda| = 0. \tag{25a}$$

Since the jump in log m is not zero, it follows that for interior impulses

$$\frac{d}{dt} \ |\lambda| = 0 \ ,$$

and (26)

$$\lambda \cdot \dot{\lambda} = 0.$$

Since λ is a continuous function with continuous derivatives (up through $\dddot{\lambda}$) then the maximum value of $|\lambda|$ is the same constant for the entire interior domain between the initial and final conditions.

For impulsive thrusts in the interior of the domain we have that $\frac{d}{dt}$ $(m\sigma) = 0$. Thus, the two new constants of the motion for the impulsive case are identical to those for the singular case within the interior domain.

b) Boundary impulses.

For a boundary impulse, $\frac{d}{dt}(m\sigma)$ vanishes only at one end of the impulse. Thus, equation (25) becomes

$$-c\log\frac{m^+}{m^-}\frac{d}{dt}|\lambda| = \begin{array}{l}\text{either } \frac{d}{dt}(m\sigma)^+ \\ \\ \text{or } -\frac{d}{dt}(m\sigma)^-\end{array} \qquad (25b)$$

The positive sign is associated with a terminating boundary impulse, and the negative sign is associated with an initiating boundary impulse. Since from equation (10)

$$\frac{d}{dt}(m\sigma) = \frac{k}{m}(|\lambda| - \frac{\sigma m}{c}) \quad ,$$

the product of an infinite, impulsive thrust and a vanishing switch function is indeterminate at an impulse. Equation (25b) may be used to evaluate this indeterminacy. At both the initial and the terminal boundary impulses, we have

$$-c\log\frac{m^+}{m^-}\frac{d}{dt}|\lambda| = \pm\frac{k}{m}(|\lambda| - \frac{\sigma m}{c}) \ . \qquad (25c)$$

In addition, from equation (5d), integrating over the boundary impulse

$$\sigma^+ - \sigma^- = c|\lambda|(\frac{1}{m^+} - \frac{1}{m^-}) \qquad (27)$$

Since at the interior boundary immediately following the impulse,

$$\sigma^+ = \frac{c|\lambda|}{m^+} \qquad (28)$$

it follows,

$$\sigma^- = \frac{c\,|\lambda|}{m^-} \ . \tag{29}$$

Thus, at the boundaries we have

$$(m\sigma)^- = (m\sigma)^+ = c\,|\lambda| \ . \tag{30}$$

The constants of the motion for impulsive thrusts at the boundaries are seen to be identical with those for interior thrusts as well as the singular case, so long as we interpret the state variables referred to their interior values at the boundary.

The natural boundary condition for minimum fuel is given by $\sigma_f = 1$. It is now possible to obtain the natural scaling factor for $|\lambda|$ from the equation

$$|\lambda| = \frac{m_f^+}{c} \ . \tag{31}$$

The initial value of σ may then be obtained from

$$\sigma_{\text{initial}} = \frac{m_f^+}{m^-_{\text{initial}}} \ . \tag{32}$$

To summarize: the general constants of the motion (which hold for all solutions) are given by

$$\begin{aligned} &\lambda \times \dot{R} + R \times \dot{\lambda} = A \ , \\ &\mu \frac{\lambda \cdot R}{r^3} + \dot{\lambda} \cdot \dot{R} - k\left(\frac{|\lambda|}{m} - \frac{\sigma}{c}\right) = h \ ; \end{aligned} \tag{33}$$

for the special case of the singular arc and for impulsive thrusts on the interior domain, the following additional constants hold:

$$\frac{1}{3}\lambda \cdot \dot{R} + \frac{2}{3} R \cdot \dot{\lambda} + c\frac{|\lambda|}{3} \log m - ht = b \ ,$$
$$m\sigma = d \ , \tag{34}$$

for the singular case $|\lambda|$ = constant,

for the impulsive case $|\lambda|_{\text{initial}} = |\lambda|_{\text{final}}$.

III. The Impulsive Solution

Given two position vectors in space, and a central angle, α, the vector velocity required to pass a free fall trajectory between the two position vectors is given by

$$\dot{R}_1 = \frac{\mu \tan \alpha/2}{p\ r_1} R_1 + \frac{p}{r_1\, r_2 \sin \alpha} (R_2 - R_1) \ . \tag{35}$$

Conversely, the velocity vector at the other end is given by

$$\dot{R}_2 = - \frac{\mu \tan \alpha/2}{p\ r_2} R_2 + \frac{p}{r_1\, r_2 \sin \alpha} (R_2 - R_1) \ . \tag{36}$$

The value of p is the magnitude of the angular momentum,

$$p = |R \times \dot{R}| = \text{constant during coast.} \tag{37}$$

This parameter may be used as a variable for the purposes of differentiating the total impulse to obtain the optimal impulsive trajectory.

Given the initial vectors R_1, $\dot{R}_1^-$ and the final vectors R_2, $\dot{R}_2^+$, it is required to find the minimum fuel necessary to go from condition one to condition two in a central force field. Let

$$\begin{aligned} \Delta V_1 &= \dot{R}_1^+ - \dot{R}_1^- \ , \\ \Delta V_2 &= \dot{R}_2^+ - \dot{R}_2^- \ . \end{aligned} \tag{38}$$

The scalar magnitudes of these impulsive changes in velocity are given by

$$\begin{aligned} \delta v_1 &= |\Delta V_1| \ , \\ \delta v_2 &= |\Delta V_2| \ . \end{aligned} \tag{39}$$

The condition for minimum fuel is

$$\frac{\partial}{\partial p}(\delta v_1 + \delta v_2) = 0 \ . \tag{40}$$

The resulting equation is given by

$$\delta v_2(\Delta V_1) \cdot \frac{\partial}{\partial p}\dot{R}_1 - \delta v_1(\Delta V_2) \cdot \frac{\partial}{\partial p}\dot{R}_2 = 0 \ . \tag{41}$$

Equation (41) is an eighth order polynomial in the variable p which may be solved by standard numerical techniques. For each real root, it is possible to evaluate the total scalar impulse and we may choose the niminum of these as our solution. The change in mass required to execute each successive impulse is given by

$$m_i^+ = m_i^- \ e^{-\delta v_i/c} \ . \tag{42}$$

IV. The Initial Conditions for the Adjoint Variables for Impulsive Thrust

The impulsive change in velocity may be obtained by integrating equation (5a)

$$\dot{R}_1^+ - \dot{R}_1^- = -c \log \frac{m_1^+}{m_1^-} \frac{\lambda_1}{|\lambda|} . \tag{43}$$

The value of $|\lambda|$ is obtained from equation (31)

$$|\lambda| = \frac{m_2^+}{c} .$$

The initial conditions for λ are

$$\lambda_1 = \frac{m_2^+}{c} \frac{\dot{R}_1^+ - \dot{R}_1^-}{\delta v_1} . \tag{44}$$

The initial value of σ is given by equation (32)

$$\sigma_1 = \frac{m_2^+}{m_1^-}$$

and is valid only for impulsive thrusts.

In order to obtain a first order approximation to the initial value of σ for the finite thrust case, resort is made to a Taylor series expansion of σm about the initial time,

$$(\sigma m)_o^+ = (\sigma m)_o^- + \frac{d}{dt}(\sigma m)_o (t - t_o) . \tag{45}$$

From equation (25b),

$$\frac{d}{dt}(\sigma m)^- = c \log \frac{m_1^+}{m_1^-} \; \frac{\lambda_1 \cdot \dot{\lambda}_1}{|\lambda_1|}$$

The value of the burning time, $t - t_o$, may be obtained from the finite, constant mass flow.

$$t - t_o = \frac{c}{k}(m_1^+ - m_1^-) \; . \tag{46}$$

The solution for the initial value of σ is given by

$$\sigma(t_o) = \frac{m_2^+}{m_1^-} + \frac{c^2}{k\, m_1^-}(m_1^+ - m_1^-) \frac{\lambda_1 \cdot \dot{\lambda}_1}{|\lambda_1|} \log \frac{m_1^+}{m_1^-} \; . \tag{47}$$

To obtain the initial value of $\dot{\lambda}$, it is necessary to obtain the variational state transition matrix. During coast, we have

$$\ddot{R} = -\mu \frac{R}{r^3} \; . \tag{48}$$

The variational equation may be written as

$$\frac{d^2}{dt^2} \frac{\partial R}{\partial \alpha} = -\frac{\mu \frac{\partial R}{\partial \alpha}}{r^3} + \frac{3\mu \; R \cdot \frac{\partial R}{\partial \alpha}}{r^5} R \; . \tag{49}$$

Let the α_i be the initial values of R and $\dot{R}$. The solution of equations (48) and (49) is the so-called variational state transition matrix, $\Phi(R, \dot{R})$. The differential equation for the adjoint variable λ is given by

$$\ddot{\lambda} = -\mu \frac{\lambda}{r^3} + 3\mu \frac{\lambda \cdot R}{r^5} R \; . \tag{50}$$

This equation is identical to equation (49). Since the initial value of Φ is the unit matrix, it follows that

$$\begin{Bmatrix} \lambda(t) \\ \dot{\lambda}(t) \end{Bmatrix} = \Phi(R, \dot{R}) \begin{Bmatrix} \lambda(t_o) \\ \dot{\lambda}(t_o) \end{Bmatrix} \tag{51}$$

The first three equations of equation (51) may be evaluated at the terminal time immediately preceding terminal thrust.

$$\lambda_2 = \left(\frac{\partial R}{\partial x_o}\right)\lambda_1 + \left(\frac{\partial R}{\partial \dot{x}_o}\right)\dot{\lambda}_1 \tag{51a}$$

Solving for $\dot{\lambda}_1$

$$\dot{\lambda}_1 = \left(\frac{\partial R}{\partial \dot{x}_o}\right)^{-1}\lambda_2 - \left(\frac{\partial R}{\partial \dot{x}_o}\right)^{-1}\left(\frac{\partial R}{\partial x_o}\right)\lambda_1 \tag{52}$$

The vector λ_2 may be obtained in a manner similar to λ_1 from the impulsive solution.

$$\lambda_2 = \frac{m_2^+}{c} \; \frac{\dot{R}_2^+ - \dot{R}_2^-}{\delta v_2} \quad . \tag{53}$$

Equation (52) is the required solution for the initial value of $\dot{\lambda}$.

The initial values of λ_1, $\dot{\lambda}_1$ and σ_1 should afford a good approximation for the iterative indirect method.

References

1. Lawden, D. F.; "Optimal Intermediate-Thrust Arcs in a Gravitational Field," Astronautica Acta, Vol. 8, No. 2, p. 106, 1962.

2. Leitmann, G.; "On a Class of Variational Problems in Rocket Flight," JASS, Vol. 26, No. 9, p. 586, 1959.

3. Melbourne, W. G.; "Three-Dimensional Optimum Thrust Trajectories for Power-Limited Propulsion Systems," ARS Journal, Vol. 31, No. 12, December 1961.

4. Breakwell, J. V.; "The Optimization of Trajectories," J. Soc. Indust. Appl. Math., Vol. 7, 215, 1959.

5. Kelley, H. J.; "Method of Gradients," Chapter 6 of Optimization Techniques, Academic Press, 1962.

6. Bryson, A. E. and Denham, W. F.; "A Steepest-Ascent Method for Solving Optimum Programming Problems," Jour. Appl. Mech. (Trans. ASME, Series E), pp. 247-257, June 1962.

20955

OPTIMUM RETRO-THRUST IN A GRAVITATIONAL FIELD

By

Carlos R. Cavoti

Space Sciences Laboratory
Missile and Space Division
General Electric Company
Box 8555, Philadelphia 1, Pa.

Report No. 1

August 30, 1963

Contract NAS 8-11040

Prepared for

National Aeronautics and Space Administration
George C. Marshall Space Flight Center
Huntsville, Alabama

OPTIMUM RETRO-THRUST IN A GRAVITATIONAL FIELD

by

Carlos R. Cavoti

Space Sciences Laboratory
Missile and Space Division
General Electric Company
Philadelphia 1, Penna.

Report No. 1

August 30, 1963

Contract NAS 8-11040

Prepared for

National Aeronautics and Space Administration
George C. Marshall Space Flight Center
Huntsville, Alabama

CONTENTS

Abstract

20955

A

The planar motion of a mass-point vehicle subject to the inverse square central gravitational attraction of a spherical planet and to a tangential retro-thrust force is considered.

It is shown that for minimum fuel consumption (free-time, free-range) problems, the control variable (mass-flow rate of the engine) may be obtained explicitly along the intermediate retro-thrust sub-arcs in terms of the state variables of the problem. The variable retro-thrust sub-arcs are shown to be integrable in closed-form and thus completely determined, except for three constants of integration. The variable retro-thrust sub-arcs are such that the magnitude of the vector velocity is constant along them.

Author

LIST OF SYMBOLS

g_o	Acceleration of gravity on the surface of the planet
h	Altitude above the surface of the planet
m	Mass of the vehicle
p	Constant Lagrange Multiplier
q	Generalized Coordinate
r_o	Radius of the planet
s	Independent variable for parametric problems
t	Time
V	Velocity
x	Curvilinear abcissa on the planet's great circle
z	Dimensionless velocity
γ	Angular position with respect to the fixed system
ζ	Independent variable variation at the terminal points
η	Dependent variable variation
θ	Angle between the vector velocity and the local horizon
λ	Dimensionless mass-flow
μ	Dimensionless mass
ν	Variable Lagrange multiplier
ξ	Dimensionless curvilinear abcissa
π	Boundary condition
ρ	Dimensionless radius

List of Symbols (cont.)

τ	Dimensionless time
ϕ	Side condition
Ω	Function to be minimized

Superscripts

$$(\ldots)' = \frac{d}{d\tau}(\ldots)$$

$$(\dot{\ldots}) = \frac{d}{ds}(\ldots)$$

Subscripts

I = Initial point

F = Final point

R = Reference value

1. Preliminary Considerations on the Variational Problem

Consider the set of equations

$$\phi_i \equiv q_i' - f_i(q_1, \ldots, q_n, \lambda, \tau) = 0 \quad , \quad i = 1, \ldots, n \ ,$$
$$\lambda_{min.} \leqq \lambda(\tau) \leqq \lambda_{max.} \quad , \quad \tau_I \leqq \tau \leqq \tau_F \ , \tag{1}$$

with the boundary conditions

$$\pi_\ell \left(q_{i_I}, q_{i_F}, \tau_I, \tau_F\right) = 0 \quad , \quad \ell = 1, \ldots, r \leqq 2n+1 \ . \tag{2}$$

Any solution of Eqs. (1) and (2) is expressed in terms of the variables

$$q_i(\tau) \ , \quad \lambda(\tau) \ , \quad \tau_I \leqq \tau \leqq \tau_F \ . \tag{3}$$

The function $\lambda(\tau)$ is called the "control variable" of the system. Since the problem has one degree of freedom associated with the control variable λ, an optimum requirement may be imposed on the solution arcs. The following variational problem of the Mayer form is therefore proposed: "Find in the class D' (*) of arcs $q_i(\tau)$, $\lambda(\tau)$, $\tau_I \leqq \tau \leqq \tau_F$, satisfying the constraints $\phi_i = 0$, $\pi_\ell = 0$, that arc which minimizes a generalized function $\Omega = \Omega\left(q_{i_I}, q_{i_F}, \tau_I, \tau_F\right)$, of the end-values."

From theory (Refs. 1 to 6) it is found that the first necessary conditions for an extremal in the class D' of arcs considered are that the Euler-Lagrange sum $\Lambda = \nu_i(\tau)\phi_i$, and the switching function $\Lambda_\lambda(\tau) = \dfrac{\partial \Lambda}{\partial \lambda}$ satisfy the equations

$$[\Lambda]_{q_i} \equiv \frac{d}{d\tau} P_i - \Lambda_{q_i} = 0 \ , \quad i = 1, \ldots, n \ , \quad P_i = \Lambda_{q_i'} \ , \tag{4}$$

(*) Arcs on which $q_i(\tau)$ is continuous while $q_i'(\tau)$ may be only piece-wise continuous in the interval (τ_I, τ_F).

$$\frac{d}{d\tau} Q - \Lambda_{\tau} = 0 \quad , \quad Q = \Lambda - q_i' P_i \quad , \tag{5}$$

and either one of the following

$$\Lambda_{\lambda}(\tau) = 0 \text{ , for admissible } \delta\lambda(\tau) \gtreqless 0 \text{ , } \tau_a \leqq \tau \leqq \tau_b \text{ , } \tag{6}$$

$$\Lambda_{\lambda}(\tau) \geqq 0 \text{ , for admissible } \delta\lambda(\tau) \geqq 0 \text{ , } \tau_a \leqq \tau \leqq \tau_b \text{ , } \tag{7}$$

$$\Lambda_{\lambda}(\tau) \leqq 0 \text{ , for admissible } \delta\lambda(\tau) \leqq 0 \text{ , } \tau_a \leqq \tau \leqq \tau_b \text{ , } \tag{8}$$

on every sub-arc $\tau_I \leqq \tau_a \leqq \tau \leqq \tau_b \leqq \tau_F$ forming the extremal arc, with a set of n non-simultaneously vanishing multipliers $\nu_i(\tau)$ continuous on every sub-arc, such that the $[(r+1) \times (2n + 2)]$ - matrix of terminal values

$$\left\| \begin{matrix} p_0 \dfrac{\partial\Omega}{\partial q_{i_I}} - P_{i_I} & p_0 \dfrac{\partial\Omega}{\partial q_{i_F}} + P_{i_F} & p_0 \dfrac{\partial\Omega}{\partial \tau_I} - Q_I & p_0 \dfrac{\partial\Omega}{\partial \tau_F} + Q_F \\ \\ \dfrac{\partial\pi_\ell}{\partial q_{i_I}} & \dfrac{\partial\pi_\ell}{\partial q_{i_F}} & \dfrac{\partial\pi_\ell}{\partial \tau_I} & \dfrac{\partial\pi_\ell}{\partial \tau_F} \end{matrix} \right\| \tag{9}$$

is of rank R < r + 1, and satisfying at junctions of sub-arcs the following Erdmann-Weierstrass vertex continuity conditions

$$\begin{aligned} P_i(\tau_c - 0) &= P_i(\tau_c + 0) \quad , \\ Q(\tau_c - 0) &= Q(\tau_c + 0) \quad , \end{aligned} \tag{10}$$

where $\tau_c - 0$ and $\tau_c + 0$ are values of the independent variable immediately before and immediately after the junction. For a normal non-singular extremal, as here assumed, the set of variable multipliers $\nu_i(\tau)$ is unique and the constant multiplier p_o in Eq. (9) may be set $p_o = 1$.

The preceding considerations have been made in order to provide the fundamentals for the problem to be treated in the following paragraphs.

2. Equations of Motion and Specific Optimality Problem

Consider the planar motion of a mass-point vehicle (m) subject to the inverse square force field of a central, spherical, non-rotating body of radius r_o; (see Fig. 1). If we assume that the thrust vector is applied in the direction of the velocity vector but in opposite sense and that the central gravitational field is expressed by $g = g_o \left(\frac{r_o}{r_o + h}\right)^2$, then the non-dimensional equations of motion of (m) in terms of the intrinsic system (orbital) with unit vectors $(\bar{u}_n , \bar{u}_t)$, are written

$$\phi_1 \equiv \xi' - \frac{z \cos\theta}{\rho} = 0 \quad , \tag{11}$$

$$\phi_2 \equiv \rho' - z \sin\theta = 0 \quad , \tag{12}$$

$$\phi_3 \equiv z' + \frac{\lambda v_e}{\mu} + \frac{\sin\theta}{\rho^2} = 0 \, , \tag{13}$$

$$\phi_4 \equiv \theta' - \left(\frac{z}{\rho} - \frac{1}{z\rho^2}\right) \cos\theta = 0 \, , \tag{14}$$

$$\phi_5 \equiv \mu' + \lambda = 0 \, . \tag{15}$$

In the previous equations,

$$\xi = \frac{x\, g_0}{V_R^{\,2}} \quad , \qquad z = \frac{V}{V_R} \quad , \qquad V_R = (g_0\, r_0)^{1/2} \, ,$$

$$\rho = \frac{r}{r_0} = \frac{r_0 + h}{r_0} \quad , \qquad \tau = \frac{t\, g_0}{V_R} \quad , \qquad \lambda = - \frac{dm/dt}{m_I\, g_0 / V_R}$$

$$\mu = \frac{m}{m_I} \quad , \qquad v_e = \frac{V_e}{V_R} \, .$$

The retro-thrust magnitude is given by $T = - ({}^{dm}/dt)\, V_e$ and thus the thrust per unit of initial weight of the vehicle assumed on the surface of the central body ($r = r_0$) is $\dfrac{T}{m_I\, g_0} = \lambda\, v_e$. It is assumed that the retro-thrust is bounded, i.e., $T_{min} \leqq T \leqq T_{max}$. Consequently, in Eqs. (11) to (15) we will take

$$\lambda_{min.} \leqq \lambda(\tau) \leqq \lambda_{max.} \quad , \qquad \tau_I \leqq \tau \leqq \tau_F \quad . \tag{16}$$

In particular it will be assumed that $\lambda_{min} = 0$. The set of Eqs. (11) to (15) is of the general form indicated in Eq. (1). Therefore, our previous considerations in paragraph 1 may be readily applied.

The variational problem to be analyzed in this paper is that of finding in the class D' of arcs $q_i(\tau)$, $\lambda(\tau)$, $i = 1, \dots, 5$, $(q_1 = \xi \, , q_2 = \rho \, , q_3 = z \, , q_4 = \theta \, , q_5 = \mu)$ satisfying

Eqs. (11) to (15) and prescribed boundary conditions of the form

$$\pi_\ell\left(q_{i_I}, q_{i_F}, \tau_I, \tau_F\right) = 0, \quad i = 1,..,5 \quad , \quad \ell = 1,..,r \leqq 11 \; ,$$

that arc which minimizes the function $\Omega = -\mu_F$.

Since $\lambda(\tau)$ is the control variable associated with the mass-flow of the engine, the previous problem may be formulated in physical terms as that of finding the optimum retro-thrust program in order to transfer the vehicle from given initial to given final conditions with minimum fuel expenditure. From the necessary conditions for an extremal analyzed in paragraph 1 and Eqs. (11) to (15) we find that the equations of the extremals are

$$\phi_1 \equiv \xi' - \frac{z\cos\theta}{\rho} = 0 \; , \tag{17}$$

$$\phi_2 \equiv \rho' - z\sin\theta = 0 \; , \tag{18}$$

$$\phi_3 \equiv z' + \frac{\lambda v_e}{\mu} + \frac{\sin\theta}{\rho^2} = 0 \; , \tag{19}$$

$$\phi_4 \equiv \theta' - \left(\frac{z}{\rho} - \frac{1}{z\rho^2}\right)\cos\theta = 0 \; , \tag{20}$$

$$\phi_5 \equiv \mu' + \lambda = 0 \,, \tag{21}$$

$$\{\Lambda\}_{\xi} \equiv \nu_1' = 0 \,, \tag{22}$$

$$\{\Lambda\}_{\rho} \equiv \nu_2' - \nu_1 \frac{z\cos\theta}{\rho^2} + \nu_3 \frac{2\sin\theta}{\rho^3} + \nu_4 \cos\theta\left(\frac{2}{z\rho^3} - \frac{z}{\rho^2}\right) = 0 \,, \tag{23}$$

$$\{\Lambda\}_{z} \equiv \nu_3' + \nu_1 \frac{\cos\theta}{\rho} + \nu_2 \sin\theta + \nu_4 \cos\theta\left(\frac{1}{\rho} + \frac{1}{z^2\rho^2}\right) = 0 \,, \tag{24}$$

$$\{\Lambda\}_{\theta} \equiv \nu_4' - \nu_1 \frac{z\sin\theta}{\rho} + \nu_2 z\cos\theta - \nu_3 \frac{\cos\theta}{\rho^2} - \nu_4 \sin\theta\left(\frac{z}{\rho} - \frac{1}{z\rho^2}\right) = 0, \tag{25}$$

$$\{\Lambda\}_{\mu} \equiv \nu_5' + \nu_3 \frac{\lambda v_e}{\mu^2} = 0 \,, \tag{26}$$

$$\Lambda_{\lambda} = \nu_3 \frac{v_e}{\mu} + \nu_5 = 0 \qquad \therefore \qquad \lambda = \lambda_{var.} \,, \tag{27}$$

$$\Lambda_{\lambda} = \nu_3 \frac{v_e}{\mu} + \nu_5 \geqq 0 \quad \therefore \quad \lambda = \lambda_{min.} \,, \tag{28}$$

$$\Lambda_{\lambda} = \nu_3 \frac{v_e}{\mu} + \nu_5 \leqq 0 \quad \therefore \quad \lambda = \lambda_{max.} \,, \tag{29}$$

$$\nu_1 \frac{z \cos\theta}{\rho} + \nu_2 z \sin\theta - \Lambda_{\lambda} \, \lambda - \nu_3 \frac{\sin\theta}{\rho^2} + \nu_4 \cos\theta \left(\frac{z}{\rho} - \frac{1}{z\rho^2} \right) = M = const. \tag{30}$$

Eq. (30) follows from Eq. (5) after considering that $\Lambda_{\tau} = 0$ and that time-independent control boundaries have been imposed. Thus, Eq. (30) holds along the extremal in the interval $\tau_I \leqq \tau \leqq \tau_F$. The latter equation is a consequence of the Euler equations and it may replace any one of them if so desired.

From theory (Refs. 1 to 6), it follows that the Weierstrass necessary condition

$$W \equiv \Lambda(\nu_i, \tau, q_i, Q_i') - \Lambda(\nu_i, \tau, q_i, q_i') - (Q_i' - q_i')\Lambda_{q_i'} =$$

$$= \left(\nu_3 \frac{v_e}{\mu} + \nu_5 \right) . \Delta\lambda \geqq 0 \,, \quad i,j = 1,..,5 \,, \tag{31}$$

must be satisfied at every element $(\nu_i , \tau , q_i , q_i') \in E$, (E = extremal) for all admissible sets $(\tau , q_i , Q_i') \neq (\tau , q_i , q_i')$, satisfying Eqs. (11) to (15). Also from theory, it follows that in our case, the Clebsch-Mayer condition

$$\Lambda_{\lambda\lambda}\, \eta_{n+1}^{2} + \Lambda_{\lambda q_j'}\, \eta_{n+1}\, \eta_j' + \Lambda_{q_j' q_i'}\, \eta_i' \eta_j' \geqq 0 , \quad i,j = 1,..,5 , \qquad (32)$$

must be satisfied at each element $(\nu_i , \tau , q_i , q_i' , \lambda) \in E$ by every set $(\eta_i' , \eta_{n+1}) \neq (0,0)$ consistent with the equations of variation

$$\Phi_i \equiv \eta_i' - \frac{\partial f_i}{\partial \lambda}\, \eta_{n+1} = 0 , \qquad i = 1,..,5 . \qquad (33)$$

Owing to the form of Eqs. (11) to (15), the Legendre condition vanishes identically. From Eqs. (27) and (31) it is also found that along λ - var. arcs the Weierstrass necessary condition is satisfied in its weak form W = 0. From the explicit expression of the Weierstrass condition and introducing

$$\omega^{\cdot} = \left(\nu_3 \frac{v_e}{\mu} + \nu_5 \right) \cdot \lambda \quad , \qquad (34)$$

it is found that the Weierstrass test may be written $W \equiv \omega - \omega^{*} \geqq 0$, (where the asterisk indicates values on E) and furthermore that, since ω

is linear in the control λ the Weierstrass condition implies that the optimum control, in the bounded interval of control (λ_{min}, λ_{max}), may be expressed by the minimality condition

$$\underset{\lambda_{min} \leqq \lambda \leqq \lambda_{max}}{Min} \left\{ \omega \left(\nu_3^*, \nu_5^*, \mu^*, \lambda \right) \right\} . \tag{35}$$

The Weierstrass condition and the minimality requirement in Eq. (35) are graphically shown in Fig. 2 assuming the strengthened form of Eqs. (28) and (29).

From Eqs. (11) to (15) it can be readily found that the extremals of our problem are non-singular since the value of the functional determinant

$$\Delta = \begin{vmatrix} \Lambda_{q_k' \nu_i} & \Lambda_{q_k' q_i'} \\ 0 & \phi_{k_{q_i'}} \end{vmatrix} , \tag{36}$$

is unity. In fact, the previous determinant has a diagonal of elements equal to unity, i.e., $\Lambda_{q_i' \nu_i} = 1$, $i = 1;2;..;5$ and $\phi_{k_{q_k'}} = 1$, $k = 1;2;..;5$, while the rest of the elements vanish. Thus, along any extremal sub-arc the slopes $q_i'(\tau)$, $i = 1,..,5$, are continuous and moreover $q_i'(\tau)$ and $\nu_i(\tau)$ have at least first order derivatives with respect to τ .

3. Explicit Form of the Optimum Retro-Thrust Program Along λ -Var. Arcs

In this paragraph we will derive a closed-form solution for the optimum variable retro-thrust program. Problems with free-time and free-range will be considered. That is, the boundary conditions are assumed such that $\frac{\partial \pi_\ell}{\partial \xi_I} = 0$ and/or $\frac{\partial \pi_\ell}{\partial \xi_F} = 0$, and $\frac{\partial \pi_\ell}{\partial \tau_I} = 0$ and/or $\frac{\partial \pi_\ell}{\partial \tau_F} = 0$.

The Transversality Condition implies that at terminal points of the extremal, the following (2n + 2) sub-conditions of transversality

$$\left(\frac{\partial \omega}{\partial q_{i_I}} - P_{i_I}\right) dq_{i_I} = 0 \;, \tag{37}$$

$$\left(\frac{\partial \omega}{\partial q_{i_F}} + P_{i_F}\right) dq_{i_F} = 0 \;, \tag{38}$$

$$\left[\frac{\partial \omega}{\partial \tau_I} - \left(\Lambda - q_i' P_i\right)_I\right] d\tau_I = 0 \;, \tag{39}$$

$$\left[\frac{\partial \omega}{\partial \tau_F} + \left(\Lambda - q_i' P_i\right)_F\right] d\tau_F = 0 \;, \tag{40}$$

must be satisfied for any set of differentials $(dq_{i_I} , dq_{i_F} , d\tau_I , d\tau_F) \neq (0,0,0,0)$ consistent with

$$\Pi_\ell \equiv \frac{\partial \pi_\ell}{\partial q_{i_I}} dq_{i_I} + \frac{\partial \pi_\ell}{\partial q_{i_F}} dq_{i_F} + \frac{\partial \pi_\ell}{\partial \tau_I} d\tau_I + \frac{\partial \pi_\ell}{\partial \tau_F} d\tau_F = 0 . \quad (41)$$

At the initial point we will take $\tau_I = 0$, $\xi_I = 0$; thus $d\tau_I = 0$, and $d\xi_I = 0$. Then, for the free final range, free final time problems, here considered, we will assume $\frac{\partial \pi_\ell}{\partial \xi_F} = \frac{\partial \pi_\ell}{\partial q_{1_F}} = 0$ and $\frac{\partial \pi_\ell}{\partial \tau_F} = 0$. Thus, dq_{1_F} $(i.e\ d\xi_F)$ and $d\tau_F$ are different from zero and may be totally arbitrary in Eqs. (41). Consequently, from Eqs. (38) and (40) it follows that

$$\frac{\partial \Omega}{\partial q_{1_F}} = -P_{1_F} \quad , \quad \frac{\partial \Omega}{\partial \tau_F} = \left(q_i' P_i - \Lambda\right)_F . \quad (42)$$

Since $\Omega = -\mu_F = -q_{5_F}$, $P_{1_F} = \nu_{1_F}$, $\left(q_i' P_i\right)_F = M = const.$, and $\Lambda = 0$, Eqs. (22), (30) and (42) lead to

$$\nu_1(\tau) = 0 \quad , \quad \tau_I \leqq \tau \leqq \tau_F \quad , \quad (43)$$

$$M = 0 = const. \quad , \quad \tau_I \leqq \tau \leqq \tau_F \quad . \quad (44)$$

Moreover, since our minimal problem is $\Omega = -\mu_F = min.$, then no boundary condition is imposed on q_{5_F} and the coefficient of dq_{5_F} in Eq. (38), vanishes. Thus, $dq_{5_F} = d\mu_F$ is different from zero and arbitrary and then Eq. (38) leads to

$$\frac{\partial \Omega}{\partial q_{5_F}} = -P_{5_F} \quad \therefore \quad \nu_{5_F} = 1 \; . \tag{45}$$

Eq. (27) leads to

$$\nu_3 \frac{v_e}{\mu} + \nu_5 = 0 \; , \tag{46}$$

and thus from Eqs. (24), (26), (43) and the total differential of Eq. (46),

$$\nu_3(\tau) = K_3 = \text{const.}, \tag{47}$$

along the $\lambda_{\text{var.}}$ sub-arc. Eq. (47) implies that

$$\nu_2 \sin\theta + \nu_4 \cos\theta \cdot \left(\frac{1}{\rho} + \frac{1}{z^2\rho^2}\right) = 0 \; . \tag{48}$$

From Eqs. (27), (30), (43) and (44),

$$\nu_2 z \sin\theta - \nu_3 \frac{\sin\theta}{\rho^2} + \nu_4 \cos\theta \left(\frac{z}{\rho} - \frac{1}{z\rho^2}\right) = 0 \; , \tag{49}$$

and then from Eqs. (48) and (49) it follows that

$$\nu_3\, z \sin\theta + 2\,\nu_4 \cos\theta = 0 \,. \tag{50}$$

Eqs. (17) to (26), and the total differential of Eq. (50), lead to

$$\nu_3\left[\left(z^2+\frac{1}{\rho}\right)\frac{\cos^2\theta}{\rho} - \left(\frac{\lambda v_e}{\mu}+\frac{\sin\theta}{\rho^2}\right)\sin\theta\right] - 2\,\nu_2\, z\cos^2\theta = 0 \,. \tag{51}$$

Consequently, the compatibility condition for a non-trivial solution $(\nu_1, \nu_2, \nu_3, \nu_4, \nu_5) \neq (0,0,0,0,0)$ is

$$\begin{vmatrix} z\sin\theta & -\dfrac{\sin\theta}{\rho^2} & \cos\theta\left(\dfrac{z}{\rho}-\dfrac{1}{z\rho^2}\right) \\ 0 & z\sin\theta & 2\cos\theta \\ -2z\cos^2\theta & \left(z^2+\dfrac{1}{\rho}\right)\dfrac{\cos^2\theta}{\rho}-\left(\dfrac{\lambda v_e}{\mu}+\dfrac{\sin\theta}{\rho^2}\right)\sin\theta & 0 \end{vmatrix} = 0 \,. \tag{52}$$

The vanishing of the previous determinant leads to the important expression

$$\lambda v_e = -\frac{\mu \sin\theta}{\rho^2} \quad , \tag{53}$$

which gives explicitly the retro-thrust program along λ-var. extremal arcs for the minimum fuel consumption problem proposed.

3. 1. Integration of the State Variables Along the $\lambda_{\text{var.}}$ Sub-arc.

An interesting physical conclusion becomes apparent once Eq. (53) is replaced in Eq. (19). In fact, after this is done we readily obtain that along $\lambda_{\text{var.}}$ sub-arcs, z = const. Thus, the variable retro-thrust program is such that the magnitude of the vector velocity is constant.

Now, from Eqs. (18), (20) and (53) the following integral may be derived

$$Ln\left(\rho\cos\theta\right) + \frac{1}{z^2\rho} = C_1 = const. \tag{54}$$

Also, from Eqs. (18), (21) and (53) we have

$$Ln\left(\mu\right) + \frac{1}{z\rho v_e} = C_2 = const. \tag{55}$$

Finally, from Eqs. (17), (18), (20) and (54) it is found that

$$\xi = \int_{\rho_I}^{\rho_F} F(\rho)\, d\rho \quad , \tag{56}$$

where

$$F(\rho) = \frac{\exp.\left(C_1 - 1/\rho z^2\right)}{\rho^2\left\{1 - \frac{\exp.\left[2\left(C_1 - 1/\rho z^2\right)\right]}{\rho^2}\right\}^{1/2}} \cdot \quad (57)$$

Thus, the variable retro-thrust sub-arc has been completely determined except for three constants of integration.

4. Integration of the Equations of Motion Along $\lambda = 0$ Sub-arcs

For $\lambda = \lambda_{\min}$ (i.e., $\lambda = 0$) the well-known integrals of the planar two-body problem are obtained. In fact, the energy integral leads to

$$\rho'^2 + \rho^2 \xi'^2 - \frac{2}{\rho} = z^2 - \frac{2}{\rho} = H = const. \; , \tag{58}$$

and from the integral of area we obtain

$$\rho^2 \xi' = \rho z \cos\theta = C_3 = const. \tag{59}$$

From Eqs. (58) and (59) it may be derived that

$$\rho = \frac{C_3}{1/C_3 + \left(H + 1/C_3^2\right)^{1/2} \cos\left(\gamma - \omega_0\right)} \tag{60}$$

which is the equation of a conic section in polar coordinates with the origin located at one of its foci. Introducing

$$\tilde{\ell} = C_3^{\,2} \quad , \quad e^2 = 1 + C_3^{\,2} H \quad , \tag{61}$$

Eq. (60) réduces to the well-known form

$$\rho = \frac{\tilde{\ell}}{1 + e \cos\left(\gamma - \omega_0\right)} \; , \tag{62}$$

where ω_0 , in the case of the ellipse, is a constant of integration determining the position of the pericentron with respect to the fixed system, $\gamma - \omega_0$ is the true anomaly (measured from the pericentron), $\tilde{\ell} = \frac{\ell}{r_0}$ is the non-dimensional semi latus rectum of the conic and e the eccentricity.

For H < 0 the conic is an ellipse, for H = 0, a parabola and for H > 0 an hyperbola. For elliptic motion

$$\tilde{\ell} = \tilde{a}\left(1 - e^2\right) , \qquad \tilde{a} = \frac{a}{r_0} , \tag{63}$$

$$H = - \frac{1}{\tilde{a}} , \quad \left(a = \text{semi major axis.}\right) , \tag{64}$$

$$C_3 = \left[\tilde{a}\left(1 - e^2\right)\right]^{1/2} . \tag{65}$$

5. Conclusions

The minimum fuel consumption problem, in the class of orbits satisfying the hypotheses made in paragraph 2, with final time and final range not specified, has been considered. This class of problems (free time, free range) gives a lower bound for the propellant needed as compared to any other solution of the problem satisfying an additional condition in the final time and/or range.

For bounded thrust magnitude (between a given maximum value and zero) the extremal arc may be composed of sub-arcs of three types: full retro-thrust, intermediate retro-thrust and zero retro-thrust (or coasting sub-arc). The sub-arcs actually forming an extremal arc, their sequence, and location of corner points has to be investigated in each specific boundary-value problem proposed. In general, these depend on the boundary conditions imposed.

The intermediate retro-thrust sub-arcs have been obtained in closed-form. The three constants of integration characterizing these sub-arcs may be determined from the values of the state variables at the end-points or at the corner points.

It has been shown that along variable retro-thrust sub-arcs the magnitude of the vector velocity is a constant.

The variable retro-thrust sub-arcs as well as the coasting sub-arcs map into points in the planes of their corresponding constants of integration. This suggests a method of piecing the extremals which will be considered as an extension of this work for given boundary-value problems.

References

1. Bliss, G. A., "Lectures on the Calculus of Variations," The University of Chicago Press, Chicago, 1946.

2. Courant, R., Hilbert, D., "Methods of Mathematical Physics," Vols. I and II, Interscience Publishers, Inc., New York, 1953 and 1962 respectively.

3. Cavoti, C. R., "Necessary and Sufficient Conditions for an Optimum in a Class of Flight Trajectories," G. E. Space Sciences Lab., T.I.S. R63SD28, March 1963.

4. Cicala, P., "An Engineering Approach to the Calculus of Variations," Libreria Editrice Universitaria Levrotto & Bella, Torino, Italy, 1957.

5. Bolza, O., "Lectures on the Calculus of Variations," Stechert-Hafner, Inc., New York, 1946.

6. Cavoti, C. R., "The Calculus of Variations Approach to Control Optimization," Special Report No. 1, Marshall Space Flight Center, Contract NAS 8-2600, Huntsville, Ala., June 1962.

7. Lawden, D. F., "Optimal Intermediate-Thrust Arcs in a Gravitational Field," Astronautica Acta, Vol. VIII, Fasc. 2-3, 1962.

8. Lawden, D. F., "Optimal Powered Arcs in an Inverse Square Law Field," ARS Journal, April 1961.

9. Kelley, H. J., "Singular Extremals in Lawden's Problem of Optimal Rocket Flight," AIAA Summer Meeting, California, 1963.

10. Leitmann, G., "On a Class of Variational Problems in Rocket Flight," Journal of the Aerospace Sciences, 26, 586, 1959.

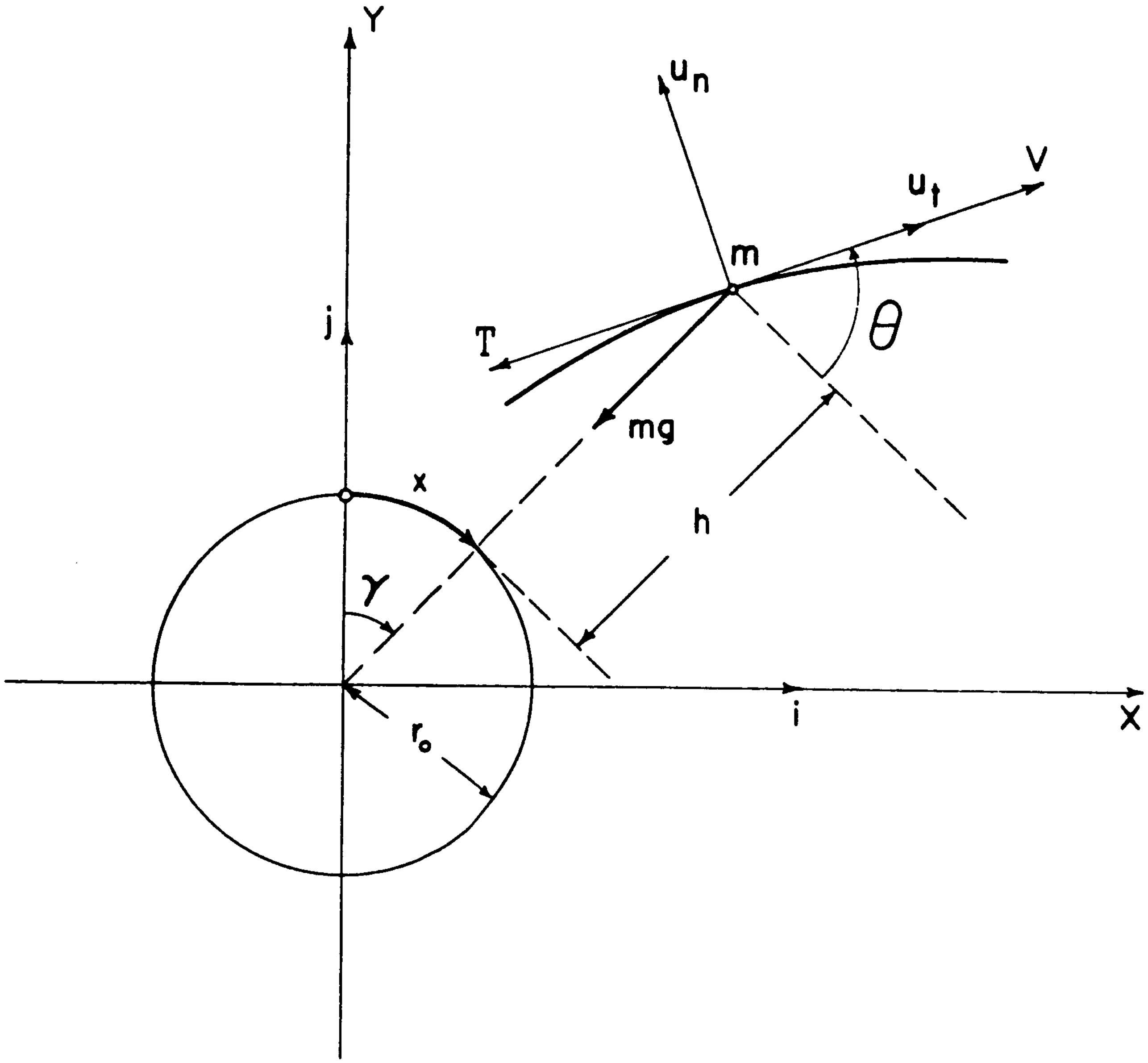
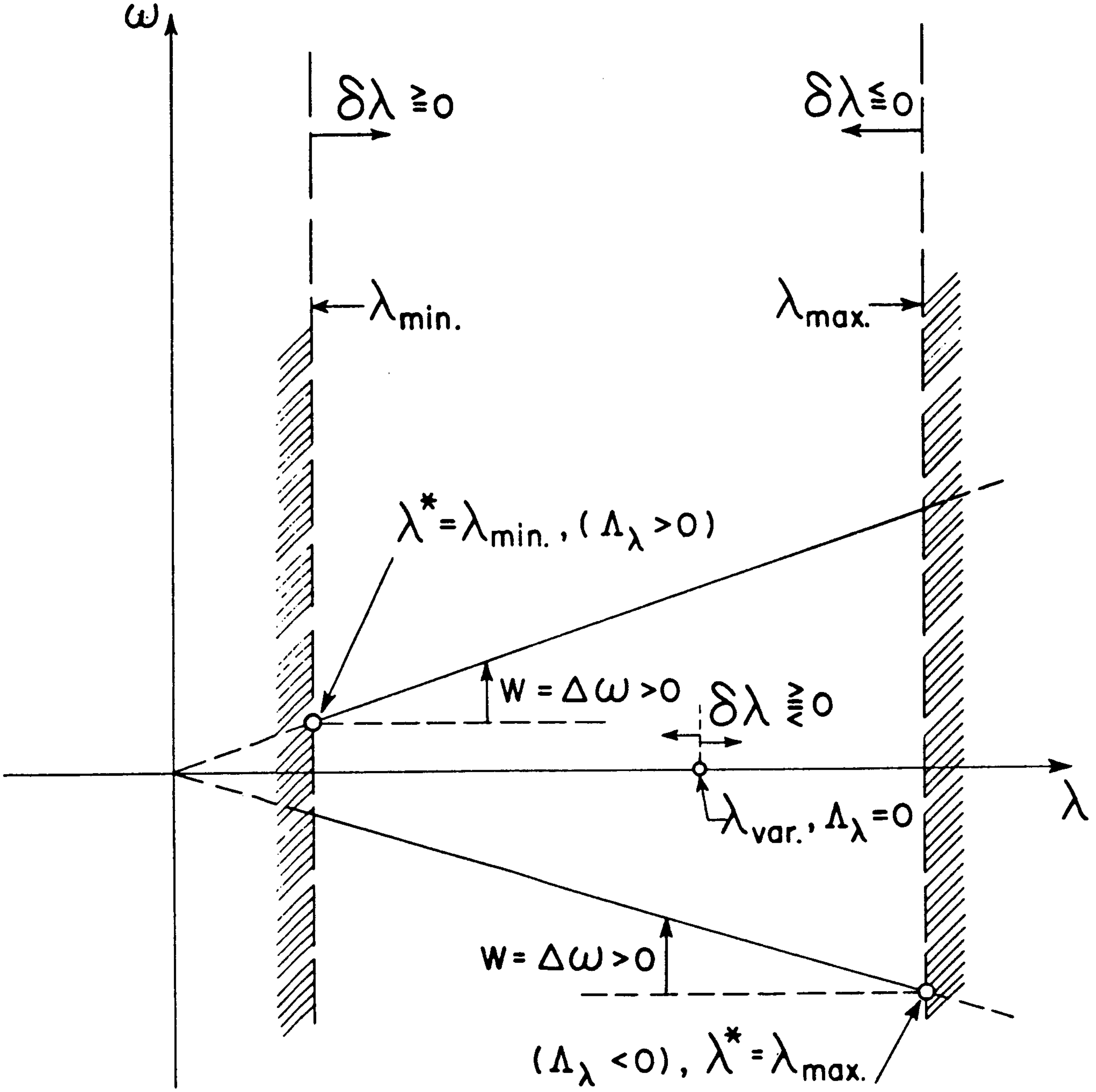


Figure 1



ω
$\delta\lambda \geqq 0$
$\delta\lambda \leqq 0$
$\lambda_{min.}$
$\lambda_{max.}$
$\lambda^* = \lambda_{min.}, (\Lambda_\lambda > 0)$
$W = \Delta\omega > 0$
$\delta\lambda \gtreqless 0$
λ
$\lambda_{var.}, \Lambda_\lambda = 0$
$W = \Delta\omega > 0$
$(\Lambda_\lambda < 0), \lambda^* = \lambda_{max.}$

Figure 2

ON A RESTRICTED COMPARISON OF TWO IMPULSE AND ONE IMPULSE ORBITAL TRANSFER

Prepared by

Gentry Lee

Space Sciences Laboratory
Space and Information Systems Division
North American Aviation, Inc.

Special Report No. 4

August 8, 1963

Contract NAS8-5211
Prepared for

George C. Marshall Space Flight Center
National Aeronautics and Space Administration
Huntsville, Alabama

ABSTRACT

20956

A

The comparison of one and two impulse orbital transfers, basic to the solution of the optimum n-impulse problem, is extended to include all copolar elliptical orbits of equal angular momentum. Familiar vector expressions are used to identify, for all cases, a specific family of two impulse transfers that require no more total impulse than the one impulse transfer applied at the intersection.

Author

I. INTRODUCTION

Two impulse orbital transfer studies have been the subject of many scientific papers(1-3) during the last few years. Recent advances, combining both the analytical(4-5) and numerical(6) viewpoints, have virtually solved the problem in its most general form. The next logical development in the general area of orbit transfers is an attempt to ascertain the degree to which the two impulse solution approximates the optimum impulse solution.

At the core of this study is a comparison between one and two impulse transfers. If Ting's suggestion(7) for copolar coplanar orbits--that there always exists a two-impulse transfer between the two orbits that requires less impulse than the best one-impulse transfer applied at the intersection--could be proven true and extended to inclined orbits, then it would naturally follow that for n arbitrarily large, the n-impulse transfer would require less fuel than any transfer using fewer impulses. This comparison, which assumes no time constraint, was carried out by Horner(8) for coplanar ellipse to circle transfers. He found that for his problem, except for notable circumstances, the optimum one-impulse transfer could always be beaten by a two-impulse transfer. Barrar[9] proved that the one-impulse transfer between intersecting orbits was always inferior to the Hohmann transfer, assuming that the orbits could be rotated to produce a Hohmann transfer. The note presented here is concerned with fixed coplanar elliptical orbits of the same angular momentum.

II. FORMULATION OF PROBLEM

Consider a plane polar coordinate system with origin at a common focus of two ellipses. Designate one orbit (the initial orbit in the transfer problem) as A (See Figure 1) and define the θ-reference line as being in the direction from the origin to the perigee of orbit A. The other orbit, designated by B, has its perigee displaced by an angle γ.

The velocity vectors of particles moving in the two orbits can be given, using familiar hodograph representation[10] and complex variables by

$$\dot{\underline{R}}_a(\theta) = \frac{h_a}{p_a} \left(1 + e_a e^{i\theta}\right) \tag{1}$$

$$\dot{\underline{R}}_b(\theta) = \frac{h_b}{p_b} \left(1 + e_b e^{i(\theta - \gamma)}\right) \tag{2}$$

Then a new function I, representing the difference between the velocity vectors at any point θ, is defined as

$$I(\theta) = \left[(\dot{\underline{R}}_a - \dot{\underline{R}}_b) \cdot (\dot{\underline{R}}_a - \dot{\underline{R}}_b)^*\right]^{\frac{1}{2}} \tag{3}$$

where the * refers to the complex conjugate. It is easily seen that if the orbits intersect for some (θ_1, θ_2), $I(\theta_1)$ and $I(\theta_2)$ represent one impulse transfers between the two orbits.

Consider a family T of transfer orbits, defined by parameters e_t, $\frac{h_t}{p_t}$, and τ. τ is the angular displacement from the reference line directed to the perigee of A. Then

$$\dot{\underline{R}}_t(\theta) = \frac{h_t}{p_t} \quad 1 + e_t \; e^{i(\theta - \tau)} \tag{4}$$

and for every triple $(e_t, \frac{h_t}{p_t}, \tau)$ where the transfer orbit intersects the initial and final orbits, there exist two impulse transfers using that particular transfer orbit.

Finally define functions M_i, $(i = 1,4)$, such that they represent two impulse transfers between A and B and such that they are defined only at those values $(e_t, \frac{h_t}{p_t}, \tau)$ where the transfer orbit intersects the initial and final orbits. Then

$$M_i(e_t, \frac{h_t}{p_t}, \tau) = \left[(\dot{\underline{R}}_a - \dot{\underline{R}}_t) \cdot (\dot{\underline{R}}_a - \dot{\underline{R}}_t)^*\right]^{\frac{1}{2}} + \left[(\dot{\underline{R}}_t - \dot{\underline{R}}_b) \cdot (\dot{\underline{R}}_t - \dot{\underline{R}}_b)^*\right]^{\frac{1}{2}} \tag{5}$$

A particular triple $(e_t{}', \frac{h_t}{p_t}, \tau')$ defines a transfer orbit T'. If that orbit intersects orbits A and B, then there are four possible ways of making a two-impulse transfer. The subscripts to the functions M can be arbitrarily determined by assigning some kind of order to the points of intersection. Thus the four combinations of points of intersection representing the possible transfers define the M_i's.

The optimum two-impulse transfer is then given by the least possible value of the M_i's. It is this value that is compared to the lower of the two values $I(\theta_1)$ and $I(\theta_2)$. That this least value of the M_i's is always less than the lower values of $I(\theta_1)$ and $I(\theta_2)$ is what Ting(7) implied was true and has been the subject of some research.

III. ORBITS OF EQUAL ANGULAR MOMENTUM

Due to the dependence of the functions M_i on the points of intersection between the orbits involved, the closed formulation of the general problem in terms of the given parameters is quite laborious. However, if orbits have the same angular momentum, an interesting property greatly simplifies the analysis.

Assume now that A and B have the same angular momentum. Then

$$\frac{h_a}{p_a} = \frac{h_b}{p_b} = \frac{h}{p} \tag{6}$$

and the function I can be written as

$$\begin{aligned} I(\theta) &= \left[(\dot{\underline{R}}_a - \dot{\underline{R}}_b) \cdot (\dot{\underline{R}}_a - \dot{\underline{R}}_b)^*\right]^{\frac{1}{2}} \\ &= \frac{h}{p}\,(e_a^2 + e_b^2 - 2e_ae_b \cos \gamma\,)^{\frac{1}{2}} \end{aligned} \tag{7}$$

The function I is now a constant, not dependent upon the angle θ. Thus for fixed copolar orbits of the same angular momentum, the difference between the velocity vectors is a constant.

The one-impulse transfer between the two orbits is easily seen to be given by equation (7). However, the formulation of the M_i's has not been considerably reduced. If, in the vector triple describing the transfer orbit, $\frac{h_t}{p_t}$ is fixed and set equal to $\frac{h}{p}$, the problem becomes susceptible to analysis. For this reduced family of transfer orbits, due to the simplification brought about by the equal angular momentum property, there exists only one M function defining the two-impulse transfers.

$$M(e_t, \tau) = \frac{h}{p} \left(e_a^2 + e_t^2 - 2e_a e_t \cos \tau\right)^{\frac{1}{2}} + \frac{h}{p} \left[e_t^2 + e_b^2 - 2e_b e_t \cos (\gamma - \tau)\right]^{\frac{1}{2}} \tag{8}$$

Now equations (7) and (8) are easily recognized as being similar to the magnitude of a side of a triangle given by the law of cosines. Define $\underline{e}_a$, $\underline{e}_b$, and $\underline{e}_t$(11) as vectors in directions of their respective perigees, having magnitudes equal to their eccentricities. From Figure 2, it is seen that M becomes only a function of the vector $\underline{e}_t$. From the law of cosines and the triangle inequality, the value of M can never be less than the value of I. However, if

$$\underline{e}_t' = \underline{e}_a + k\,(\underline{e}_b - \underline{e}_a), \qquad 0 \leq k \leq 1 \tag{9}$$

then it is obvious that the value for M (the two impulse transfer) is equal to the value of I. There exists, therefore, a family of two impulse transfers that give the same impulse as the one impulse transfer. Only if $\underline{e}_a = \underline{e}_b$, in which case the initial and final orbits would be coincident, would this family vanish.

The nature of the orbits for this family of two-impulse transfers can be deduced from qualitative reasoning. Subsequent algebraic investigation of the intersections of these orbits proved this reasoning to be accurate. Since the angular momentum is not changed in these transfers, the only component of the velocity that is altered is the radial component. This fact, when combined with another property of copolar ellipses of the same angular momentum (they intersect 180 degrees apart), leads to the realization that this family represents a splitting of the impulse at the intersection point of A and B. Each orbit with angular momentum equal to the angular momentum of A and B and with $\underline{e}_t'$ defined by (8) passes through both intersection points of A and B. In the two impulse transfers that equal the one impulse, a certain percentage of the radial velocity change is used at the first intersection point and then, after a 180 degree coast, the final velocity change injects into orbit B.

IV. CONCLUSIONS

It has been shown here that for copolar elliptical orbits of the same angular momentum, there exists a specific family of two impulse transfers that use no more impulse than the one impulse transfer at the intersection. If it is true for every $\underline{e}_a$, $\underline{e}_b$, and $\frac{h}{p}$ combination that there exists at least one k, $0 \leq k \leq 1$, such that the total impulse as a function of the angular momentum is not a minimum at $\frac{h_t}{p_t} = \frac{h}{p}$, then Ting's suggestion would hold true for orbits of the same angular momentum. However, a general study, by means of vector analysis, of the quantities involved was unable to produce this proof. Indeed, recent intensive numerical investigations[(12)] have shown that there are many orbital configurations for which the optimum two impulse and one impulse transfers vary by only a slight amount. Three years have shown that the general proof of Ting's statement is elusive; perhaps more investigation in this area will be able to demonstrate further regions of validity (or invalidity) of his suggestion.

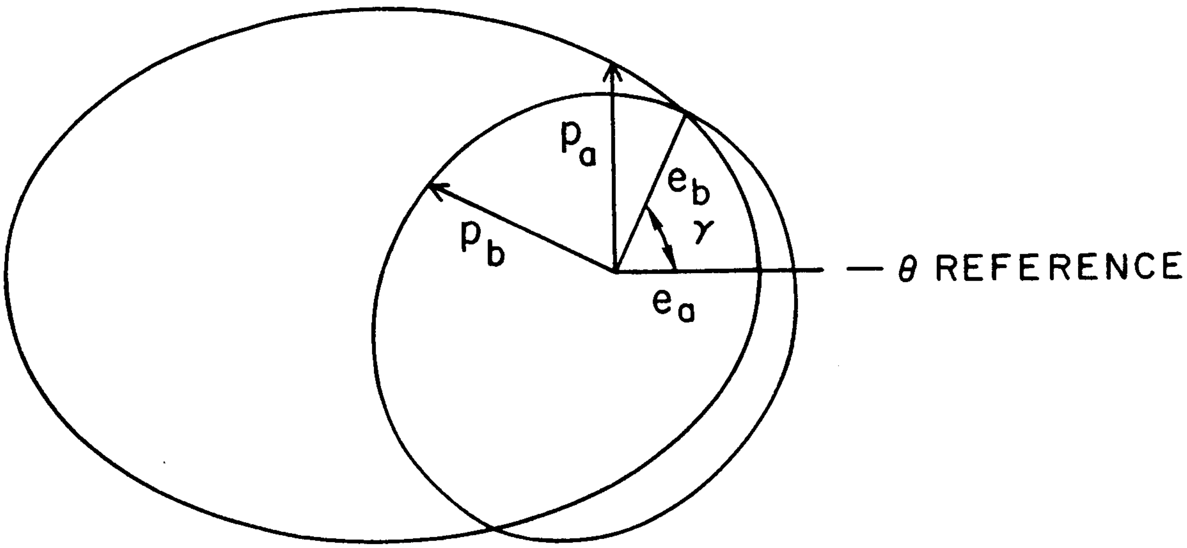


FIGURE I. GEOMETRY OF COPOLAR ELLIPSES OF EQUAL P

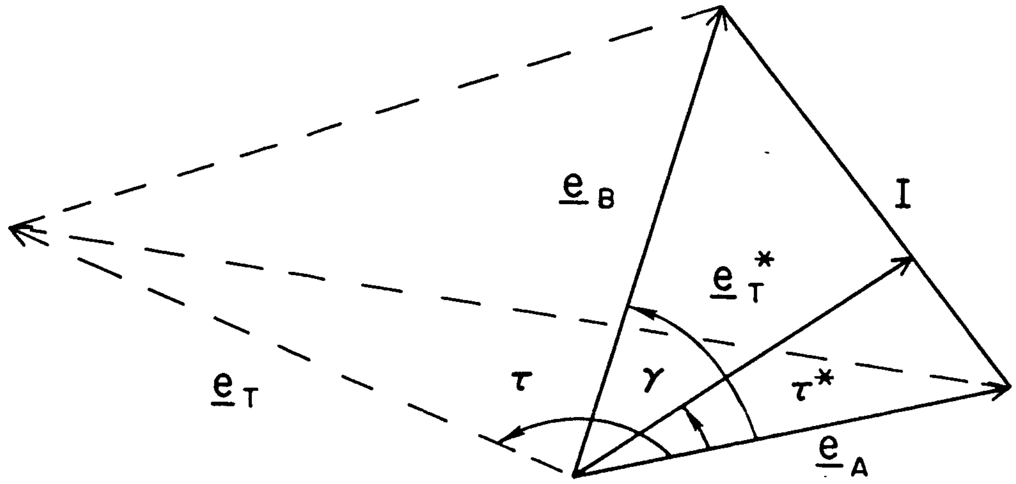


FIGURE 2. TRIANGLE INEQUALITY PROVING $M \geq I$ FOR $\frac{h_t}{p_t} = \frac{h}{p}$

REFERENCES

1. Lawden, D. F., "Optimal Two-Impulse Transfer," Optimization Techniques, (Academic Press, New York, edited by George Leitmann), 333-348, (1962).

2. Horner, J. M., "Optimum Two-Impulse Transfer Between Arbitrary Coplanar Terminals," ARS Journal, Vol. 32, 95-96 (1962).

3. Altman, S. P. and Pistiner, J. S., "Minimum Velocity Increment Solution for Two-Impulse Coplanar Orbital Transfer," AIAA Journal, Vol. 1, 435-442, (1963).

4. Altman, S. P. and Pistiner, J. S., "Analysis of Orbital Transfer Problem in Three Dimensional Space," presented at AIAA Astrodynamics Conference, New Haven, Conn., August 19, 1963.

5. Lee, Gentry, "An Analysis of Two-Impulse Orbital Transfer," Space Sciences Laboratory, North American Aviation, Inc., SID 63-741, July 1, 1963. (To be published elsewhere.)

6. McCue, G. A., "Optimum Two-Impulse Orbital Transfer and Rendezvous Between Inclined Elliptical Orbits," AIAA Journal, Vol. 1, 1865-72 (1963).

7. Ting, L., "Optimum Orbital Transfer by Impulses," ARS Journal, Vol. 30, 1013-1018 (1960).

8. Horner, J. M., "Optimum Impulsive Orbital Transfers Between Coplanar Orbits," ARS Journal, Vol. 32, 1082-1089 (1962).

9. Barrar, R. B., "Two-Impulse Transfer vs. One-Impulse Transfer: Analytic Theory," AIAA Journal, Vol. 1, 65-69 (1963).

10. Altman, S. P. and Pistiner, J. S., "Hodograph Transformations and Mapping of the Orbital Conics," ARS Journal, Vol. 32, 1109-1111 (1962).

11. Herget, Paul, "The Computation of Orbits," Published privately by author, Ann Arbor, Michigan, 1948, p. 30.

12. McCue, G. A., "On Two Impulse vs. One Impulse Orbital Transfer," Private communication, August 1963 (To be published).

A COMPARISON OF ONE AND TWO IMPULSE TRANSFER FOR NEARLY TANGENT COPLANAR ELLIPTICAL ORBITS

Prepared by

D. F. Bender

Space Sciences Laboratory
Space and Information Systems Division
North American Aviation, Inc.

Special Report No. 5

December 16, 1963

Contract NAS8-5211
Prepared for

George C. Marshall Space Flight Center
National Aeronautics and Space Administration
Huntsville, Alabama

20957

SUMMARY

A

It is proved, given two elliptic orbits that are tangent at a point not an apsis, and given that a pair of near by intersections is obtained by sufficiently small variations of the elements (from the tangent condition), that the impulse for transfer at one of the intersections will be less than that for transfer at tangency.

For the case that the intersections are caused and deepened by changing only the relative orientation, numerical results show the one impulse transfer to pass through a minimum. On the same graph there is also plotted the impulse for optimum two-impulse transfer and it is seen that near the one impulse minimum there is a region where one-and two-impulse transfers require the same impulse to two parts in 10^5 for the moderately eccentric cases chosen. It is suggested that this type of behavior may be the fundamental reason why two-impulse transfer is so close to optimum impulsive transfer even if it is not really the optimum.

Author

LIST OF SYMBOLS

e, p, ω	Orbital elements; eccentricity, semi-latus rectum, argument of perigee (orbit 2 with respect to orbit 1)
C	Given by $C^2 = \rho^2 e_1{}^2 - 2\rho\, e_1 e_2 \cos\omega + e_2{}^2$
D	Given by $D^2 = \rho^4 e_1 - 2\rho^2\, e_1 e_2 \cos\omega + e_2{}^2$
E	Given by $D^2 = \rho^3 e_1{}^2 - (\rho + 1)\rho\, e_1 e_2 \cos\omega + e_2{}^2$
j	Impulse to transfer in unit of $\sqrt{\frac{\mu}{p_1}}$
$\underline{Q}$	Unit vector perpendicular to perigee direction
r	Radius
$\underline{V}$	Unit vector perpendicular to radius
ϵ	Half angle between the two intersections
μ	Gravitation constant (G times mass)
ϕ	Perigee of orbit 1 measured from direction bisecting the angle between the intersections (arbitrary when introduced)
$\rho \sqrt{p_2/p_1}$	

I. INTRODUCTION

The study of impulsive transfer between two given coplanar elliptic orbits around an attracting body has so far yielded no proofs showing whether one, two, or more impulses yield the minimum impulsive transfer. There are special cases in which three-impulses are better than two,[1] but for the most part the best two-impulse transfer between two given orbits seems to be practically unbeatable. It was noticed very early in studies at the Space Sciences Laboratory[2] that the optimum transfer orbits were very nearly tangent at both departure and arrival points to the given orbits. This is suggested also by the fact that co-tangential transfer[3,4] is a good approximation to optimum two-impulse transfer over a wide range of orbit shapes. However it is known that, given two elliptical orbits which are tangent at a point which is not an apsis, the optimum two-impulse transfer is slightly better than the single impulse transfer at the point of tangency. This can be proved by an analysis similar to that below and it is evident from the numerical results presented.

Now given the optimum two impulse transfer between the two originally given orbits one must ask whether or not the single impulse transfers utilized at either the departure or arrival points could be replaced by a two-impulse transfer that would require less total impulse. If the answer is affirmative, then three impulse transfer is better than two, 4 better than 3, etc. However if a better two-impulse transfer cannot be found to replace either of the transfers at departure and arrival then perhaps the optimum two-impulse transfer is in fact the optimum impulsive transfer. This paper does not purport to answer this question definitively. However, it is demonstrated that over a narrow but finite range of orbit shapes for shallowly intersecting orbits one-impulse and optimum-two impulse transfer require practically identical total impulses. Thus we are led to investigate the properties of a single impulse transfer for two orbits that intersect shallowly. Such pairs of intersections are obtained by applying small variation(s) to the elements of one or the other of the two orbits starting from a tangency condition that is not an apsis. The theorem which will be proved is that one of the two intersections requires less impulse than the tangency situation and the other requires more as the intersection is initiated. Further the one requiring less impulse is expected to pass through a minimum as the intersection deepens. The fact that the tangent case is not an optimum single impulse transfer for fixed shape orbits that may have arbitrary relative perigees was shown by L. Ting[5].

In the last portion of this note numerical results are presented showing, for the case in which the intersection is produced by rotating the two orbits, the nature of the one impulse minimum. For this range of shapes the impulse required for the best two-impulse transfer is also indicated, and it is seen that the two curves are extremely close together over a range of shapes near the minimum of the one impulse curve. The fact that this region is finite in width may make it possible for the transfer orbit for the optimum two-impulse case to satisfy such a condition at both ends. This may be the fundamental reason why two-impulse transfer is so close to the optimum impulsive transfer even if it is not really the optimum.

II. THE SHALLOW INTERSECTION

Consider two elliptical orbits that are nearly tangent. These are characterized by the elements p_1, $e_1 \neq 0$, $\omega_1 = 0$ and $p_2 = \rho^2 p_1$, $e_2 \neq 0$, $\omega_2 = \omega$. The angular reference direction is taken as perigee of the first orbit and thus ω is simply the difference in the two perigee directions. We require $\omega \neq 0$, $e_1 \neq 0$, $e_2 \neq 0$ in order that the tangent case be not an apsis. Thus circular orbits are excluded from the discussion. For $p_1 = p_2$, (or $\rho = 1$) the intersections of the orbit (if any exist) must lie 180° apart and we exclude this case because a shallow intersection is to be characterized by a small angle between the two points of intersection.

In order to determine the points of intersection of two coplanar orbits we first introduce an arbitrary reference direction so that $\omega_1 = \emptyset$, $\omega_2 = \emptyset + \omega$. The angles are illustrated in Figure 1. Let one of the intersections be at ϵ. Equating the two expressions for radius at this point gives

$$r = \frac{p_1}{1 + e_1 \cos(\epsilon - \emptyset)} = \frac{p_2}{1 + e_2 \cos(\epsilon - \emptyset - \omega)} \tag{2.1}$$

Thus

$$\begin{aligned} p_1 + p_1 e_2 \cos\epsilon \;\; \cos(\emptyset + \omega) + p_1 e_2 \sin\epsilon \;\; \sin(\emptyset + \omega) \\ = p_2 + p_2 e_1 \cos\epsilon \;\; \cos\emptyset + p_2 e_1 \sin\epsilon \;\; \sin\emptyset \end{aligned} \tag{2.2}$$

Now we choose $\emptyset$ so that the terms involving $\sin\epsilon$ cancel, thus requiring

$$\tan\emptyset = \frac{p_1 e_2 \sin\omega}{p_2 e_1 - p_1 e_2 \cos\omega} = \frac{e_2 \sin\omega}{\rho^2 e_1 - e_2 \cos\omega} \tag{2.3}$$

where $\rho^2 = \dfrac{p_2}{p_1}$

The value of ϵ is obtained from

$$\cos\epsilon = \frac{p_1 - p_2}{p_2 e_1 \cos \phi - p_1 e_2 \cos(\phi + \omega)} = \frac{1 - \rho^2}{\rho^2 e_1 \cos \phi - e_2 \cos(\phi + \omega)} \quad . \tag{2.4}$$

By using

$$\sin \phi = \frac{e_2 \sin\omega}{D} \tag{2.5}$$

$$\cos \phi = \frac{\rho^2 e_1 - e_2 \cos\omega}{D} \tag{2.6}$$

We find

$$D^2 = \rho^4 e_1{}^2 - 2\rho^2 e_1 e_2 \cos\omega + e_2{}^2 \tag{2.7}$$

and

$$\cos\epsilon = \frac{\rho^2 - 1}{D} \quad . \tag{2.8}$$

Since there are two intersections it is clear that they must lie at ± ϵ and the reference direction must bisect the angle between the two intersections. Suppose now that the orbits are tangent to one another and that this is indicated by the subscript (T). This requires $\cos\epsilon_T = \pm 1$ and the two values of ϵ_T are either ± 0, or ± 180°. Which pair it is depends on the quadrant chosen for ϕ or on the sign of D. We suppose that ϕ and D are chosen so that the tangent case will be $\epsilon_T = \pm 0$, as indicated in Figure 1. Equation (2.8) yields

$$D_T = \rho_T{}^2 - 1 \tag{2.9}$$

We are concerned with small variations in the elements p_1, p_2, e_1, e_2 and ω from the values at tangency and since only the ratio p_2/p_1 is involved we have replaced p_1 and p_2 with $\rho\ (=\sqrt{p_2/p_1})$. Thus $\rho = \rho_T + \delta\rho$, $e_1 = e_{1T} + \delta e_1$, etc., where $\delta\rho$, δe_1, etc. are the small changes.

We assume $\epsilon << 1$ and we may write

$$1 - \cos\epsilon \simeq \frac{\epsilon^2}{2} \simeq - \sum_{j=1}^{4} \frac{\partial \cos\epsilon}{\partial \alpha_j} \delta\alpha_j \tag{2.12}$$

where α_j are the four elements: ρ, e_1, e_2, and ω. (Note that D involves the elements, α_j, and its derivatives are included.) Thus

$$\epsilon = \pm \sqrt{-2 \sum_{j=1}^{4} \frac{\partial(\cos\epsilon)}{\partial \alpha_j} \delta\alpha_j} \ . \tag{2.13}$$

In the numerical comparisons below the shallow pair of intersections will be generated by rotating one of the two tangent orbits with respect to the other. Thus ρ, e_1, e_2 are considered fixed and only ω is changed in the proper direction. We find

$$D^2 - D_T^2 = 2\rho^2 e_1 e_2 \left[\cos\omega_T - \cos(\omega_T + \delta\omega)\right] \simeq 2\rho^2 e_1 e_2 \sin\omega \quad \delta\omega$$

and

$$\epsilon \simeq \frac{\rho}{\rho^2 - 1} \sqrt{2 e_1 e_2 \sin\omega \ (\delta\omega)} \ . \tag{2.14}$$

In the coefficient of $\delta\omega$ no distinction is made between ω_T and ω.

III. ONE IMPULSE TRANSFER AT SHALLOW INTERSECTIONS

The velocities at the point of transfer are expressed in units of $\sqrt{\mu/p_1}$. Thus

$$\frac{\underline{v}_2}{\sqrt{\frac{\mu}{p_1}}} = \underline{V} + e_1 \underline{Q}_1 \tag{3.1}$$

$$\frac{v_2}{\sqrt{\frac{\mu}{p_1}}} = \frac{1}{\rho}\,(\underline{V} + e_2\underline{Q}_2) \tag{3.2}$$

$\underline{V}$ is the unit vector perpendicular to the radius at the transfer point and $\underline{Q}_1$, $\underline{Q}_2$ are unit vectors perpendicular to the perigee directions (See Fig. 1). The impulse to make the transfer, in units of $\sqrt{\mu/p_1}$, is expressed as

$$\underline{j} = \frac{\underline{V} + e_2\underline{Q}_2 - \rho\underline{V} - \rho e_1\underline{Q}_1}{\rho} \tag{3.3}$$

$$= \frac{\underline{V}\,(1-\rho) + e_2\underline{Q}_2 - \rho\, e_1\underline{Q}_1}{\rho} \quad = \frac{\underline{V}\,(1-\rho) + \underline{C}}{\rho}$$

where

$$\underline{C} = e_2 Q_2 - \rho\, e_1\, \underline{Q}_1 \tag{3.4}$$

Then

$$j^2\,\rho^2 = C^2 + (1-\rho)^2 + 2(1-\rho)\;\underline{C}\cdot\underline{V} \tag{3.5}$$

where

$$C^2 = \rho^2 e_1{}^2 - 2\rho\, e_1 e_2 \cos\omega + e_2{}^2 \tag{3.6}$$

and

$$\underline{C}\cdot\underline{V} = e_2\,\underline{Q}_2\cdot\underline{V} - \rho\, e_1\,\underline{Q}_1\cdot\underline{V} \tag{3.7}$$

$$= e_2 \cos\,(\epsilon - \phi - \omega) - \rho e_1 \cos\,(\epsilon - \phi)$$

The angles $(\epsilon - \phi - \omega)$ and $(\epsilon - \phi)$ are the true anomalies of the transfer point on the second and the first orbits respectively. By using Eq. (2.5) and (2.6) the angle ϕ is eliminated and Eq. (3.7) reduces to:

$$\underline{C}\cdot\underline{V} = + \frac{\cos\epsilon}{D}\;E^2 + \frac{\sin\epsilon}{D}\;(1-\rho)\rho\;\; e_1 e_2 \sin\omega \tag{3.8}$$

where

$$E^2 = \rho^3 e_1{}^2 - (1 + \rho)\ \rho e_1 e_2 \cos\omega + e_2{}^2 \qquad (3.9)$$

Thus

$$j^2 \rho^2 = C^2 + (\rho - 1)^2 - 2(\rho - 1)\left[\frac{E^2}{D}\cos\epsilon + \frac{(1 - \rho)\rho e_1 e_2 \sin\omega \sin\epsilon}{D}\right] \qquad (3.10)$$

Now we can make the comparison of the impulse for the tangent condition j_T with those for the two points of intersection: j_{x1} and j_{x2}. The only difference between these two cases is the sign of ϵ and so only one expression (Eq. 3.10) is needed. For the tangent case

$$j_T{}^2 = \frac{1}{\rho_T{}^2} C_T{}^2 + \frac{(\rho_T - 1)^2}{\rho_T{}^2} - \frac{2(\rho_T - 1)E_T{}^2}{\rho_T{}^2 D_T} \qquad (3.11)$$

Again we assume ϵ to be small and consider only small changes (infinitesimal) in ρ, e_1, e_2 and ω.

Since we may write

$$j_{x1}{}^2 - j_T{}^2 = (j_{x1} - j_T)(j_{x1} + j_T) \simeq (j_{x1} - j_T)\, 2j_T \qquad (3.12)$$

we find

$$j_{x1} - j_T \simeq \frac{1}{2j_T}\left[\left(\frac{C^2}{\rho^2} - \frac{C_T{}^2}{\rho_T{}^2}\right) + \left(\frac{(\rho - 1)^2}{\rho^2} - \frac{(\rho_T - 1)^2}{\rho_T{}^2}\right)\right. \qquad (3.13)$$

$$\left. + \left(-\frac{2(\rho-1)E^2}{\rho^2 D}\cos\epsilon + \frac{2(\rho_T - 1)E_T{}^2}{\rho_T{}^2 D_T}\right) + \frac{2(\rho-1)^2 e_1 e_2 \sin\omega}{\rho D}\epsilon\right]$$

Now all the paired terms in Eq. (3.1) can be expressed as Taylor series about the tangent condition and the leading terms involve $\delta\rho$, δe_1, δe_2 or $\delta\omega$. The term involving ϵ however has in its leading term $\sqrt{\delta\rho}\sqrt{\delta e_1}$,

$\sqrt{\delta e_2}$, or $\sqrt{\delta\omega}$ from the expression for ϵ. (Eq. 2.13). Therefore as long as the ϵ term does not have a zero coefficient it will dominate the expression at first as the small changes are added. Since ϵ can be positive or negative it follows that for one intersection the impulse to transfer is at first less than that required at tangency, and for other intersections it is greater.

It is believed that the restriction allowing only changes that cause the tangency to yield a pair of real intersections will make all of the paired terms in Eq. (3.12) positive. This is the case for changes in ω (Eq. (3.14) below), but it has not been proved for changes in ρ, e_1, or e_2. However if it is true, than a minimum one-impulse case will occur as the intersection is deepened by a continuous change in any of the elements unless higher order terms interfere.

For the case that ρ, e_1, e_2 are fixed and only ω is varied, Eq. (3.13) yields:

$$j_{x1} - j_T \simeq \frac{e_1 e_2 \sin\omega}{j_T} \left[\delta\omega \frac{2\,E^2}{(\rho + 1)\,D^2} + \frac{(\rho - 1)^2}{\rho\,D}\,\epsilon\right] \qquad (3.14)$$

Removing ϵ by using Eq. (2.14) gives for ϵ neg.

$$j_{x1} - j_T \simeq \frac{e_1 e_2 \sin\omega}{j_T\,(\rho + 1)} \left[\delta\omega\, 2\,\frac{E^2}{D^2} - \sqrt{\delta\omega}\,\frac{\sqrt{2 e_1 e_2 \sin\omega}}{\rho + 1}\right] \qquad (3.15)$$

The terms neglected in Eq. (3.15) begin with $\delta\omega^{3/2}$, and $\delta\omega$ has to be positive in the direction which yields the pair of shallow intersections.

Since the sign of the coefficient of $\delta\omega$ is positive Eq. (3.15) has a minimum which is given by

$$(\delta\omega)_m = \frac{e_1 e_2 \sin\omega}{8(\rho + 1)^2\, E^4/D^4} \qquad (3.16)$$

The corresponding values of ϵ and $j_{x1} - j_T$ are

$$\epsilon_m = \frac{\rho}{\rho^2-1} \; \frac{e_1 e_2 \sin\omega}{2(\rho + 1) \; E^2/D^2} \tag{3.17}$$

and

$$(j_{x1} - j_T)_m = - \frac{e^2{}_1 e^2{}_2 \sin^2\omega}{4 j_T(\rho + 1)^3 \; E^2/D^2} \tag{3.18}$$

IV. NUMERICAL COMPARISON OF ONE AND TWO-IMPULSE TRANSFERS NEAR THE TANGENT CASE

A program for one-impulse transfers was developed by G. A. McCue who supplied the one-impulse data presented. In addition several two-impulse optimum transfers for the cases considered were supplied by G. A. McCue who utilized the program described in his report on Optimum Two Impulse Orbital Transfer[6].

Two orbit pairs were selected for the study. They are: case (1) $\rho^2 = 1.2$, $e_1 = e_2 = .2$; and case (2) $\rho^2 = 1.8$, $e_1 = .2$, $e_2 = .6$. The corresponding values of ω for tangency are: (1) $\omega_T = \cos^{-1} .6 = 53^\circ\!.1301$ and (2) $\omega_T = 110^\circ\!.3741$. In Table 1 there are collected the values of the constants and the values of $\delta\omega_m$, ϵ_m, $(j_T - j_{x1})_m$ for both cases.

The values indicated "(pred.)" were obtained from Eqs. (3.16), (3.17), and (3.18) while the values labeled "(comp.)" were obtained by the one impulse computer program. It can be seen that the predicted values are quite close to the actual values obtained and the equations do give, for the case of an orbit pair rotated to tangency and then to shallow intersections, the approximate size and shape of the one impulse transfer versus perigee angle curve.

A program for obtaining the best 180° two-impulse transfer 4 was used as a guide in testing results because of its simplicity. In fact the data used for the two-impulse curves shown in Figures 2 and 3 were obtained with this program. Points on this curve obtained by the two-impulse optimization program are indicated by black dots. They are indeed at a lower total impulse than the 180° curve but on the scale shown the difference is not significant. The investigation of the real nature of these small differences is a subject for further work.

In both cases shown the one and two impulse curves agree to within 2 parts in 10^5 over a finite range of relative orientations and hence of relative shape. In this region there is no practical advantage as far as total velocity change is concerned whether one or two-impulse transfer is used.

TABLE 1. PARAMETERS CONCERNING ONE IMPULSE TRANSFER NEAR TANGENCY

	CASE 1	CASE 2
Fixed elements	$\rho^2 = 1.2$, $e_1 = e_2 = .2$	$\rho^2 = 1.8$, $e_1 = .2$, $e_2 = .6$
Perigee difference for tangent case (deg)	53.1301	110.3741
Impulse at tangency unit $\sqrt{\mu / p_1}$	.0853686 (1971.31)*	.296964 (6259.9)**
$(\delta\omega)_m$ (pred.)	.059	.174
(deg) (comp.)	.060	.172
$(\epsilon)_m$ (pred.)	2.56	2.51
(deg) (comp.)	2.56	2.60
$(j_x - j_T)_m$ (pred.)	-.000348	-.00087
(comp.)	-.000348 (8.03)*	-.00089 (18.9)**

* In ft/sec for $p_1 = 5000$ miles

** In ft/sec for $p_1 = 6000$ miles

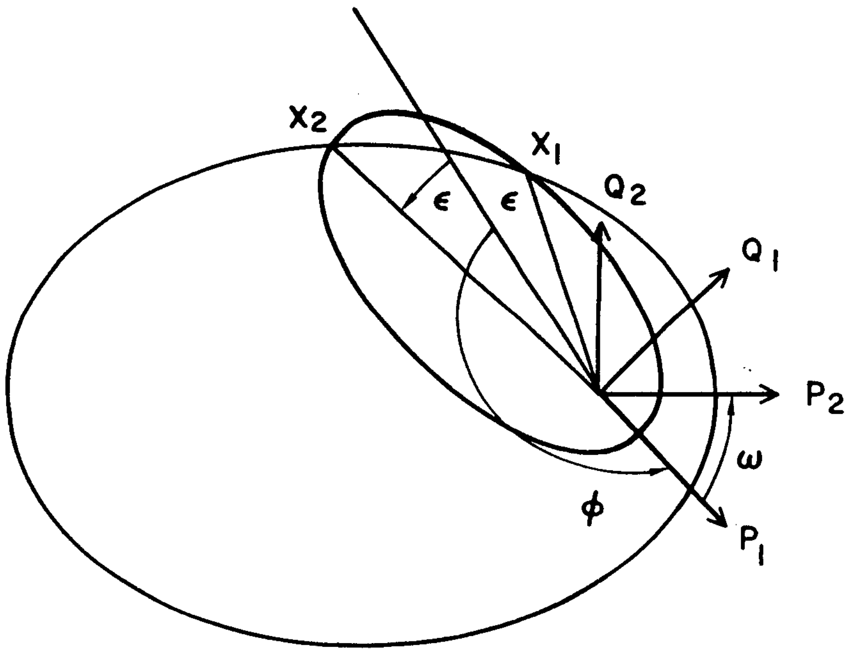


FIG. 1. THE GEOMETRY OF SHALLOW INTERSECTIONS

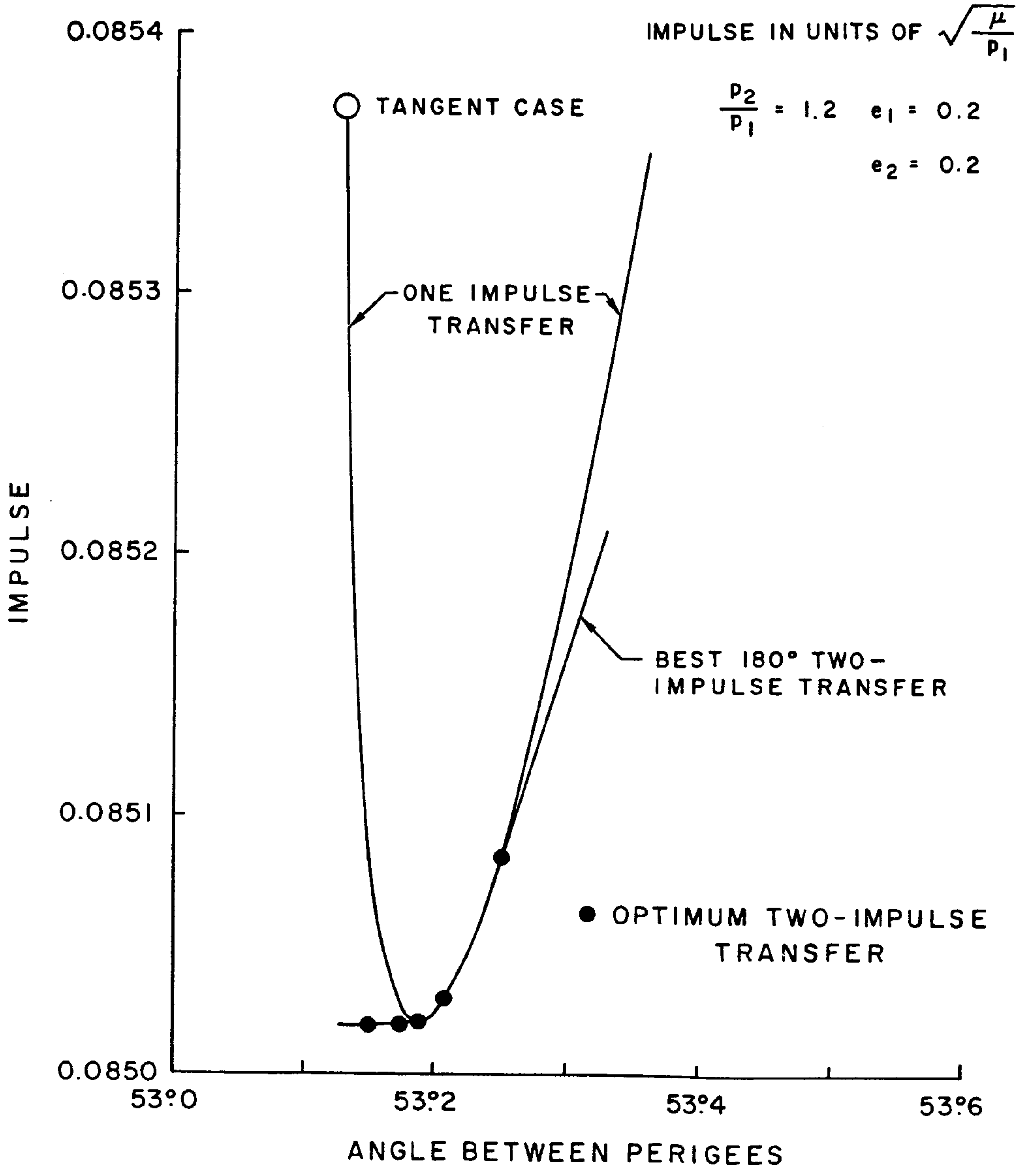


FIG. 2. COMPARISON OF ONE AND TWO-IMPULSE TRANSFERS FOR NEARLY TANGENT ORBITS

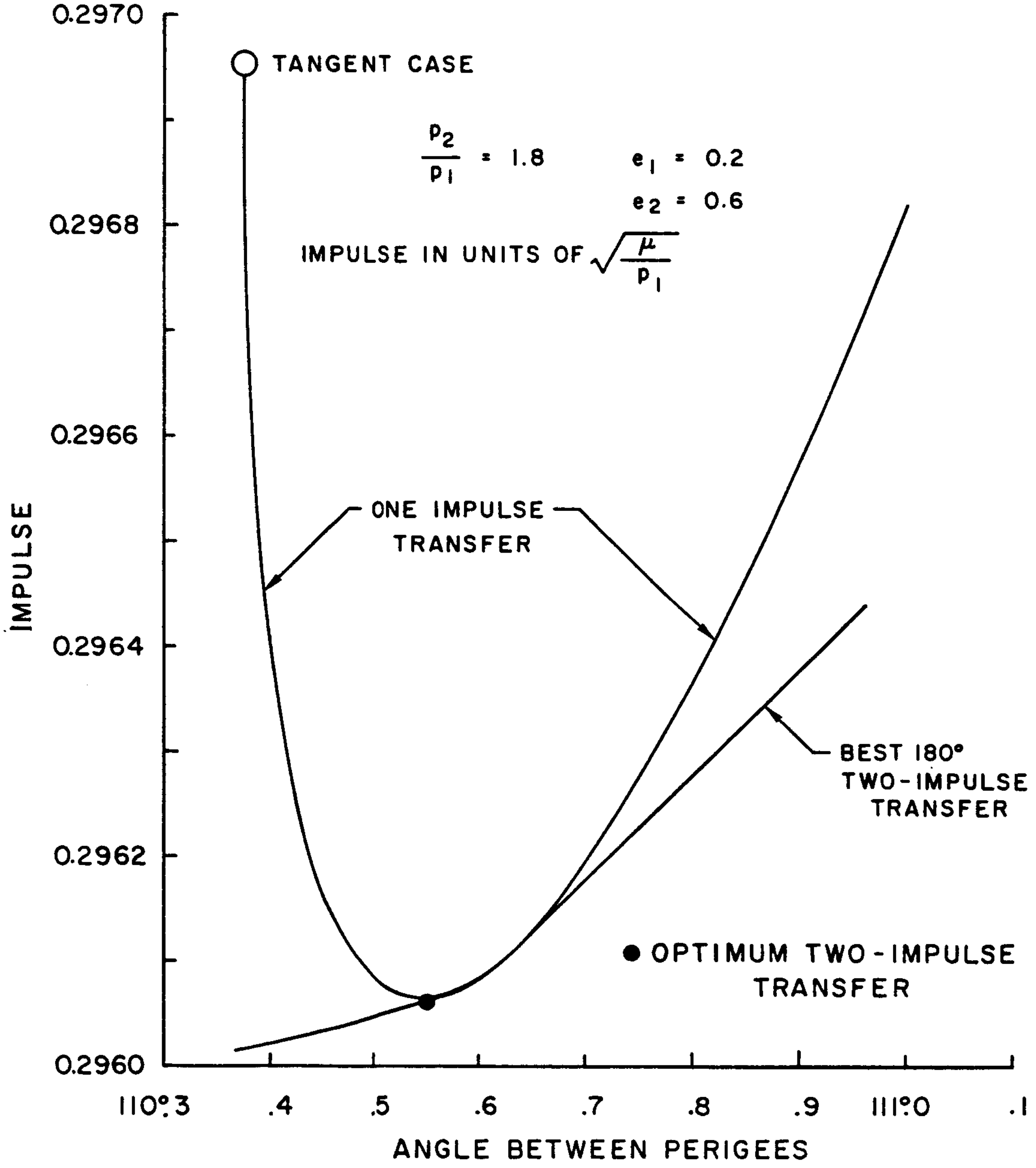


FIG. 3 COMPARISON OF ONE AND TWO-IMPULSE TRANSFERS FOR NEARLY TANGENT ORBITS

REFERENCES

1. Hoelker, R. and Silber, R., "The Bi-elliptical Transfer Between Circular Coplanar Orbits DA.TM, 2-59, ABMA Redstone Arsenal, Alabama (Jan. 1959).

2. Kerfoot, H. P. and DesJardins, P. R., "Coplanar Two Impulse Orbital Transfers," ARS preprint 2063-61 (October 9-15, 1961) and "Analytical Study of Satellite Rendezvous, Final Report", MD 59-272, Space and Information Systems Division, North American Aviation, Downey, California.

3. Lawden, D. F., "Orbital Transfer via Tangential Ellipses," Journal of the British Interplanetary Society, 11, 278-89 (1952).

4. Bender, D. F., "Optimum Coplanar Two-Impulse Transfer Between Elliptic Orbits," J. of Aerospace Eng., 21, 44-52 (1962).

5. Ting, L., "Optimum Orbital Transfer by Impulse," ARS Journal, 30, 1013-18 (1960).

6. McCue, G. A., "Optimum Two-Impulse Transfer and Rendezvous between Inclined Elliptic Orbits," AIAA Journal 1, 1865-72 (1963), and Progress Report No. 4 on Studies in the Fields of Space Flight and Guidance Theory, pages 135-165.

Department of Mathematics and Astronomy
University of Kentucky
Lexington, Kentucky

A Matrix Representation of the General Cubic and its Translation

by

The Kentucky Team
T. J. Pignani, H. G. Robertson, J. B. Wells, Jr,
and J. C. Eaves, Director

Special Report No. 7
18 December 1963
Contract NAS 8-1611

Prepared for
George C. Marshall Space Flight Center
Huntsville, Alabama

Department of Mathematics and Astronomy
University of Kentucky
Lexington, Kentucky

A Matrix Representation of the General Cubic and its Translation

by

The Kentucky Team

T. J. Pignani, H. G. Robertson, J. B. Wells, Jr., and

J. C. Eaves, Director

A

30958

SUMMARY

The purpose of this report is to present a matrix method for representing the general cubic $\sum_{i,j,k} \alpha_{ijk} x_i x_j x_k$ and to give directly the coefficients of the cubic subjected to the transformation $x_i = y_i + \beta_i$, $i = 1, 2, \cdots, n$. This method enables one to compute the coefficients of the new cubic form in any order and to apply approximation in the final summing stages.

A MATRIX REPRESENTATION OF THE GENERAL CUBIC AND ITS TRANSLATION

The matrix representation of quadratic forms X^TAX is well known, [1], where $X^T = (x_1x_2 \cdots x_n)$, and A is the $n \times n$ matrix (a_{ij}). The algebra of matrices and its application to space missile theory is available in excellent form in a number of available publications, [2], [3]. Higher dimensional matrices mentioned in the mathematical literature occasionally but mostly as an introduction to the study of tensor algebra [4]. In this paper we use the usual laws of matrix algebra and the extended associative law for multiplication of $n \times n \times n$ matrices by vectors, X, X^T, and X^D (the depth element) and, with this, represent the general cubic $\sum_{i,j,k}^{n} a_{ijk}x_ix_jx_k$ in the form $X^T/X^D/AX$. It is convenient to denote the $n \times n \times n$ matrix A in such a way that

$$\overset{\daleth}{A} = \begin{pmatrix} a_{i11} & a_{i12} & \cdots & a_{i1n} \\ a_{i21} & a_{i22} & \cdots & a_{i2n} \\ \cdots & & & \\ a_{in1} & a_{in2} & \cdots & a_{inn} \end{pmatrix}$$

represents the ith slice of A. In.general the element a_{ijk} is in the ith slice jth row and the kth column. We have for

$$X = \begin{pmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{pmatrix}, \quad X^T = \begin{pmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{pmatrix}^{T} \quad X^D = \begin{pmatrix} & & & & x_n \\ & & & \cdot & \\ & & \cdot & & \\ & x_2 & & & \\ x_1 & & & & \end{pmatrix}.$$

The definition of the product $X^D A$ is given by

$$X^D A = \sum_{i=1}^{n} x_i A^{\overline{i}\rceil} .$$

The ith summand of $X^D A$ is $x_i A^{\overline{i}\rceil}$. Thus, $X^T/X^D/AC$ is given by $\sum_{i=1}^{n} x_i X^T A^{\overline{i}\rceil} X$. From this it follows that, in the general cubic, the coefficient of

$$x_i^3 \text{ is } a_{111},\ x_2^3 \text{ is } a_{222},\ \ldots,\ x_n^3 \text{ is } a_{nnn}.$$

Partial coefficients of the $x_i^2 x_j$ term appear for $x_i^2 x_j$, for $x_i x_j x_i$, and for $x_j x_i^2$, and are a_{iij}, a_{iji}, and a_{uii} respectively. We note that a_{iij} and a_{iji} are the ij and ji elements of the ith slice of A and a_{jii} is the diagonal element of the ith row, ith column of the jth slice. Selecting $a_{iij} = a_{iji} = a_{jii}$ is in line with preserving symmetry. The term $x_i x_j x_k$ occurs with the partial coéfficients a_{ijk}, a_{ikj}, a_{jik}, a_{jki}, a_{kij}, and a_{kji} and choosing these equal preserves symmetry in the $n \times n \times n$ matrix A. As a check we note that A contains n^3 elements and that n of these are associated with the x_i^3, $i = 1, 2, \cdots, n$, that $3n(n-1)$ are associated with the $x_i^2 x_j$, $i \neq j$, and $n(n-1)(n-2)$ with the $x_i x_j x_k$, $(i, j, k, \neq)$

Translation from x_i reference to y_i where $X = Y+\beta$, $x_i = y_i+\beta_i$ is given by

$$X^T/X^D/AX = \sum_{i=1}^{n} x_i X^T A^{\overline{i}\rceil} X = \sum_{i=1}^{n} (y_i+\beta_i)(Y+\beta)^T A^{\overline{i}\rceil}(Y+\beta),$$

with $\beta^T = (\beta_k \beta_2 \cdots \beta_n)$. Using the distributive law and matrix multiplication for these matrices, we have

$$\sum_{i=1}^{n} (y_i Y^T A^{\overline{i}\rceil} Y + \beta_i Y^T A^{\overline{i}\rceil} Y + y_i \beta^T A^{\overline{i}\rceil} Y + \beta_i \beta^T A^{\overline{i}\rceil} Y +$$
$$y_i Y^T A^{\overline{i}\rceil} \beta + \beta_i Y^T A^{\overline{i}\rceil} \beta + y_i \beta^T A^{\overline{i}\rceil} \beta + \beta_i \beta^T A^{\overline{i}\rceil} \beta).$$

An examination of this sum yields the coefficients of the $y_i y_j y_k$:

y_i^3 has the coefficient a_{iii}, as expected,

$y_i^2 y_i$ has the coefficient $(a_{iij} + a_{iji} + a_{jii})$, $i \neq j$,

$y_i y_j y_k$ has the coefficient $(a_{ijk} + a_{ikj} + a_{jik} + a_{jki} + a_{kij} + a_{kji})$, $(i, j, k, \neq)$.

y_i^2 has the coefficient $\sum_{t=1}^{n} \beta_t \alpha_{tii} + \sum_{t=1}^{n} \beta_t \alpha_{iti}$

$+ \sum_{t=1}^{n} \alpha_{iit}\beta_t$ which simplifies to

$$\sum_{t=1}^{n} \beta_t(\alpha_{tii} + \alpha_{iti} + \alpha_{iit}).$$

$y_i y_j$ (for $i \neq j$) has the coefficient $\sum_{t=1}^{n} \beta_t(\alpha_{tij} + \alpha_{tji}) +$

$\sum_{t=1}^{n} \beta_t(\alpha_{itj} + \alpha_{jti}) \quad + \quad \sum_{t=1}^{n} \beta_t(\alpha_{ijt} + \alpha_{jit})$

which may be written more compactly as

$$\sum_{t=1}^{n} \beta_t \alpha_{p(i,j,t)}$$

where $p(i, j, t)$ means the summation is over all permutations of i, j, t for given i, j.

The coefficient of y_i is $\sum_{t=1}^{n} (\beta_t \sum_{s=1}^{n} \beta_s \alpha_{tsi}) + \sum_{t=1}^{n} (\beta_t \sum_{s=1}^{n} \alpha_{tis}\beta_s) +$ $\sum_{t=1}^{n} \sum_{s=1}^{n} \beta_s \alpha_{ist}\beta_t$, which may be written more compactly as

$$\sum_{t=1}^{n} \sum_{s=1}^{n} \{\beta_t\beta_s(\alpha_{tsi} + \alpha_{tis} + \alpha_{ist})\} .$$

The constant term, as expected, is $\sum_{t=1}^{n} \beta_t \beta^T A^{\overline{i}} \beta$, that is, the original function $X^T/X^D/AX$ with x_i replaced by β_i.

To illustrate, we find that the coefficient of

y_5^3 is α_{555},

$y_5^2 y_7$ is $\alpha_{557} + \alpha_{575} + \alpha_{755}$,

$y_5 y_7 y_9$ is $\alpha_{579} + \alpha_{597} + \alpha_{759} + \alpha_{795} + \alpha_{957} + \alpha_{975}$,

y_5^2 is $\sum_{t=1}^{n} \beta_t(\alpha_{t55} + \alpha_{5t5} + \alpha_{55t})$,

$y_5 y_7$ is $\sum_{t=1}^{n} \beta_t \alpha_{p(5,t,7)}$

y_5 is $\sum_{t=1}^{n} \sum_{s=1}^{n} \{\beta_t \beta_s (\alpha_{ts5} + \alpha_{t5s} + \alpha_{5st})$

Transformations of the type used to test matrices for property "p" [5], [6] suggest that a more judicious selection of the elements of the coefficient matrix of the general cubic will simplify the coefficients of the translated cubic and yet preserve the symmetry of the $\overline{A^i}$.

Select the α_{ijk} in such a way that

αx_i^3 implies $\alpha = \alpha_{iii}$

$\alpha x_i^2 x_j$ implies $\alpha = \begin{cases} \alpha_{iji} + \alpha_{iij} = 2\alpha_{iji},\ \alpha_{jii} = 0, & \text{for } i < j \\ \alpha_{jii},\ \alpha_{iji} = \alpha_{iij} = 0 & \text{for } i > j. \end{cases}$

$$\alpha x_i x_j x_k \text{ implies } \alpha = \alpha_{ijk} + \alpha_{ikj} = 2\alpha_{ijk},$$

$$\alpha_{jik} = \alpha_{jki} = \alpha_{kij} = \alpha_{kji} = 0, \text{ for } i < j,\ k.$$

Examination of the modified n×n×n matrix A reveals that it is somewhat like the *completely triangular matrix* in skeleton [7]. In fact it is *completely pyramidal* in that A^1 may contain nonzero elements in every ij position, A^i contains only zero elements in the first (i-1) rows and (i-1) columns, ..., A^n contains only one possible nonzero element, α_{nnn}. It is of interest to note that the coefficients of the x_i^3 occur in the dimensional diagonal $\alpha_{111}, \alpha_{222}, \cdots, \alpha_{nnn}$. Also we note that

n elements are associated with the x_i^3,

$2n(n-1) - \frac{(n-1)n}{2}$ elements are associated with the $x_i^2 x_j$,

$2\binom{n}{3}$ elements are associated with the $x_i x_j x_k$.

From this we see that

$$n + \{2n(n-1) - \frac{(n-1)n}{2}\} + 2\binom{n}{3} = \frac{n(2n+1)(n+1)}{6} = \sum_{i=1}^{n} i^2$$

which verifies the pyramidal skeleton of A.

In this case A is called /(1, 1, 1); (n, n, n)/ pyramidal where the $(1, 1, 1)$ gives the only possible nonzero element in the first slice and the possible nonzero base is the nth slice.

If computations are necessary, it is desirable to reverse the assignments of the coefficients a_{ijk} and obtain A in the $/(n, n, n), (1, 1, 1)/$ pyramidal form. This permits shorter summations and the use of Cyclic Entry-Exit Programming techniques which require shorter computing time [8]. It is also of interest to note that for $x_n = 1$, $X^T/X^D/AX$ is the general nonhomogeneous cubic in $x_1, x_2, \cdots, x_{n-1}$ which thus contains all cubic, quadractic and linear terms, and the constant term a_{nnn}. Also, the familiar trilinear form is given by $X^T Y^D AZ$.

REFERENCES

1. Parker, W. V. and Eaves, J. C., Matrices, The Ronald Company, 1960 .

2. Miner, W. E., Methods for Trajectory Computation, George C. Marshall Space Flight Center, January, 1963.

3. Faddeev, D. K. and Faddeeva, V. N., Computational Methods of Linear Algebra, translated by R. C. Williams, W. H. Freeman and Co., 1963.

   Gantmacher, F. R., Applications of the Theory of Matrices, New York, Interseience Publishers, 1959.

4. Mal'cev, A. I., Foundations of Linear Algebra, translated by T. C. Brown, W. H. Freeman and Co., 1963.

5. Eaves, J. C., On set of matrices having a delayed commutative commutativity property, J. of Elisha Mitchell So. vol. 68, No. 1, June 1952, p. 46-54.

6. McCoy, N. H., On Quasi Commutative matrices, Tran. A. Math So., vol. 36, p. 327-340, 1934.

7. Eaves, J. C., A note on sets of matrices simultaneoulsy reducible to the triangular skeleton, J. of Math and Physics vol. 32, no. 4, Jan. 1954, p. 302-306.

8. Eaves, J. C. and Pignani, T. J., Cyclic Entry-Exit for Subroutines, Trans. Ky. Aca. of Science, vol. 21, p. 3-4 1960.

COMPUTATION CENTER
UNIVERSITY OF NORTH CAROLINA

THE APPLICATION OF LINEAR PROGRAMMING TECHNIQUES TO RATIONAL APPROXIMATION PROBLEMS, II

By

Shigemichi Suzuki

CHAPEL HILL, NORTH CAROLINA

COMPUTATION CENTER
UNIVERSITY OF NORTH CAROLINA
CHAPEL HILL, NORTH CAROLINA

THE APPLICATION OF LINEAR PROGRAMMING TECHNIQUES TO RATIONAL APPROXIMATION PROBLEMS, II

By

Shigemichi Suzuki

SUMMARY

20959

A

This paper formulates a number of linear programming algorithms for approximating, in the sense of Tchebycheff, a function of many variables by a ratio of linear forms over a finite point set. It is a continuation of the work reported on in Progress Report No. 4, MSFC Report No. MTP-AERO-63-65, September 19, 1963. Author

§ 1

This paper describes new iterative algorithms for obtaining the "best" rational approximation of "fixed form" to a function of many variables whose value is known at a given set of data points. These methods are offered as alternatives to the algorithm developed for solving problem (a) in [1] and may be more effective than the previous method of solution. In the new algorithms, auxiliary functions are optimized under the previously given constraints. This approach makes available new information which can be used to determine bounds for ϵ^* , the optimum solution for the original non-linear program. These bounds are, in turn, used to compute successive values for ϵ in the algorithms.

As in [1], problem (a) is formulated as follows:

$f(\vec{z})$ is a function whose value is known at n points, $\vec{z}_1, \ldots, \vec{z}_n$, in a multidimensional space. $\{P_i\}_{i=1}^M$ and $\{Q_j\}_{j=1}^N$ are known functions of $\vec{z}$. Define

$$(1) \qquad R(\vec{z}) = \frac{\sum_{i=0}^M A_i P_i(\vec{z})}{\sum_{j=0}^N B_j Q_j(\vec{z})} = \frac{P(\vec{z})}{Q(\vec{z})}$$

with M and N fixed and A_i and B_j unknown. The problem is that of minimizing

$$\underset{1 \le k \le n}{\mathrm{Max}} |f(\vec{z}_k) - R(\vec{z}_k)| \ .$$

The associated non-linear problem is:

Minimize ϵ

subject to the constraints

$$|f(\vec{z}_k) - R(\vec{z}_k)| \le \epsilon \qquad (k = 1, \ldots, n) \quad . \tag{2}$$

Each constraint of (2) can be represented by the pair of inequalities

$$\begin{cases} f(\vec{z}_k) - R(\vec{z}_k) \le \epsilon \\ -\, f(\vec{z}_k) + R(\vec{z}_k) \le \epsilon \end{cases} \qquad (k = 1, \ldots, n) \quad . \tag{3}$$

Assuming that $R(\vec{z})$ does not have a pole on the set of points $\{\vec{z}_k\}_{k=1}^{n}$ and that $P(\vec{z})$ and $Q(\vec{z})$ do not have a common factor, then $Q(\vec{z}_k) > 0$ or $Q(\vec{z}_k) < 0$ for each k . We will assume that $Q(\vec{z}_k) > 0$ for all k . Hence $Q(\vec{z}_k) \ge c > 0$, $(k = 1, \ldots, n)$, for some positive number c . Now (3) becomes:

$$\begin{cases} f(\vec{z}_k)Q(\vec{z}_k) - P(\vec{z}_k) \le Q(\vec{z}_k)\epsilon \\ -\, f(\vec{z}_k)Q(\vec{z}_k) + P(\vec{z}_k) \le Q(\vec{z}_k)\epsilon \\ \qquad\qquad -\, Q(\vec{z}_k) \le -c \end{cases} \qquad (k = 1, \ldots, n) \quad . \tag{4}$$

Substituting (1) into (4), the problem becomes:

Minimize ϵ

subject to the constraints

$$\begin{cases} -\, \Sigma_{i=0}^{M} p_{ki}A_i + \Sigma_{j=0}^{N} y_k q_{kj} B_j - \Sigma_{j=0}^{N} q_{kj} B_j \epsilon \le 0 \\ \Sigma_{i=0}^{M} p_{ki}A_i - \Sigma_{j=0}^{N} y_k q_{kj} B_j - \Sigma_{j=0}^{N} q_{kj} B_j \epsilon \le 0 \\ \qquad\qquad\qquad -\, \Sigma_{j=0}^{N} q_{kj} B_j \le -c \end{cases} \qquad (k = 1, \ldots, n) \quad , \tag{5}$$

where $p_{ki} = P_i(\vec{z}_k)$, $q_{kj} = Q_j(\vec{z}_k)$, and $y_k = f(\vec{z}_k)$.

The following notation is introduced to simplify the discussion.

$$(6)\quad \begin{cases} g_k(A,B,\epsilon) = -\sum_{i=0}^{M} p_{ki}A_i + \sum_{j=0}^{N} y_k q_{kj} B_j - \sum_{j=0}^{N} q_{kj} B_j \epsilon \\ g_{n+k}(A,B,\epsilon) = \sum_{i=0}^{M} p_{ki}A_i - \sum_{j=0}^{N} y_k q_{kj} B_j - \sum_{j=0}^{N} q_{kj} B_j \epsilon \qquad (k = 1,2,\ldots,n) \\ g_{2n+k}(B) = \sum_{j=0}^{N} q_{kj} B_j \end{cases}$$

Program (5) now becomes

$$(7)\quad \begin{cases} \text{Minimize } \epsilon \\ \text{subject to the constraints,} \\ \qquad g_k(A,B,\epsilon) \le 0 \\ \qquad g_{n+k}(A,B,\epsilon) \le 0 \qquad (k = 1,2,\ldots,n) \\ \qquad -g_{2n+k}(B) \le -c \end{cases}$$

If ϵ is assigned some positive value, the constraints of program (7) become linear in the unknowns A_i and B_j. There is no objective function associated with the linearized constraints. Here, as in [1], ϵ is considered to be a parameter. By iterating on ϵ , the optimum ϵ^* can be reached as closely as desired. Upper bounds for ϵ^* as a function of the parameter ϵ are obtained through solving an associated linear program, and from these bounds a new value of ϵ is computed. After solving the linear program with a particular value of ϵ , a rational function is obtained which can be used as an approximation for the best rational function.

§ 2

By considering the following modification of program (7), an algorithm for iterating on ϵ can be constructed. The successive values of the new objective function, λ , give bounds for ϵ^* and the iterant ϵ .

$$(8)\quad \begin{cases} \text{Minimize } \lambda \\ \text{subject to the constraints} \\ \qquad g_k(A,B,\epsilon) \le \lambda \\ \qquad g_{n+k}(A,B,\epsilon) \le \lambda \qquad (k = 1,2,\ldots,n) . \\ \qquad -g_{2n+k}(B) \le -c \end{cases}$$

Some properties of programs (7) and (8) will first be given. These properties will then be used to develop the new algorithms for obtaining the optimum solution to the original problem (5). Let $\lambda^*(\epsilon)$ denote the optimum λ of program (8) and ϵ^* denote the optimum ϵ of program (7).

Property 1

There is a feasible solution for program (8) if and only if there exists a vector B such that $g_{2n+k}(B) \ge c \quad (k = 1,2,\ldots,n)$.

Proof: If there exists a vector B such that $-g_{2n+k}(B) \le -c$, then there exist an A , ϵ , and λ such that $g_k(A,B,\epsilon) \le \lambda$ and $g_{n+k}(A,B,\epsilon) \le \lambda$, $(k = 1,2,\ldots,n)$, since A , ϵ and λ are unrestricted. Hence program (8) is feasible. The necessity is obvious.

Property 2

$\lambda^*(\epsilon)$ is a strictly decreasing function of ϵ for $\epsilon \le \epsilon^*$.

Proof: Let $(A^*(\epsilon_0), B^*(\epsilon_0), \lambda^*(\epsilon_0))$ be an optimum solution of program (8) with $\epsilon = \epsilon_0$. For $\Delta\epsilon > 0$,

$$g_k(A^*(\epsilon_0), B^*(\epsilon_0), \epsilon_0 + \Delta\epsilon)$$
$$= g_k(A^*(\epsilon_0), B^*(\epsilon_0), \epsilon_0) - g_{2n+k}(B^*(\epsilon_0)) \cdot \Delta\epsilon \leq \lambda^*(\epsilon_0) - c \cdot \Delta\epsilon, \quad \text{and}$$

$$g_{n+k}(A^*(\epsilon_0), B^*(\epsilon_0), \epsilon_0 + \Delta\epsilon)$$
$$= g_{n+k}(A^*(\epsilon_0), B^*(\epsilon_0), \epsilon_0) - g_{2n+k}(B^*(\epsilon_0)) \cdot \Delta\epsilon \leq \lambda^*(\epsilon_0) - c \cdot \Delta\epsilon$$
$$(k = 1,2,\ldots,n) \quad .$$

Hence $(A^*(\epsilon_0), B^*(\epsilon_0), \lambda^*(\epsilon_0) - c\Delta\epsilon)$ is a feasible solution of program (8) with $\epsilon = \epsilon_0 + \Delta\epsilon$. Therefore

$$\lambda^*(\epsilon_0 + \Delta\epsilon) \leq \lambda^*(\epsilon_0) - c \cdot \Delta\epsilon \; .$$

Hence $\lambda^*(\epsilon_0 + \Delta\epsilon) < \lambda^*(\epsilon_0)$.

Property 3

(i) $\lambda^*(\epsilon) > 0$ if and only if $\epsilon < \epsilon^*$.

(ii) $\lambda^*(\epsilon) = 0$ if and only if $\epsilon = \epsilon^*$.

(iii) $\lambda^*(\epsilon)$ is $-\infty$ if and only if $\epsilon > \epsilon^*$.

Proof:

(i) If $\epsilon < \epsilon^*$, then for any A and B , $g_t(A,B,\epsilon) > 0$ for some t , $1 \leq t \leq 2n$. Therefore $\lambda^*(\epsilon) > 0$ if $\epsilon < \epsilon^*$.

Suppose $\lambda^*(\epsilon) > 0$. Then the constraints of program (7) are not satisfied and hence $\epsilon < \epsilon^*$. Therefore if $\lambda^*(\epsilon) > 0$, $\epsilon < \epsilon^*$.

(ii) Suppose $\lambda^*(\epsilon) = 0$. The constraints of program (7) are satisfied and $\epsilon^* \leq \epsilon$. If $\epsilon^* < \epsilon$ then by property 2, $\lambda^*(\epsilon^*) > 0$. This implies by (i) that $\epsilon^* < \epsilon^*$. Therefore $\epsilon = \epsilon^*$.

Suppose $\epsilon = \epsilon^*$. Then $\lambda^*(\epsilon) \leq 0$. If $\lambda^*(\epsilon) < 0$, and $(A^*(\epsilon)$, $B^*(\epsilon)$, $\lambda^*(\epsilon))$ is an optimum solution of program (8), then

$$g_k(A^*(\epsilon),\ B^*(\epsilon),\ \epsilon - \Delta\epsilon) \le 0$$

$$g_{n+k}(A^*(\epsilon),\ B^*(\epsilon),\ \epsilon - \Delta\epsilon) \le 0 \qquad (k = 1,2,\ldots,n)$$

$$-g_{2n+k}(B^*(\epsilon)) \le -c$$

whenever $\Delta\epsilon \le -\dfrac{\lambda^*(\epsilon)}{c}$.

This implies that program (7) is feasible with $\epsilon = \epsilon^* - \Delta\epsilon$. This contradicts the definition of ϵ^* . Therefore $\lambda^*(\epsilon) = 0$.

(iii) Suppose $\lambda^*(\epsilon)$ is $-\infty$. By (i) and (ii), $\epsilon > \epsilon^*$.

Suppose $\epsilon > \epsilon^*$. Let $(A^*(\epsilon^*),\ B^*(\epsilon^*),\ \lambda^*(\epsilon^*))$ be an optimum solution of program (8) with $\epsilon = \epsilon^{*\prime}$. By (ii), $\lambda^*(\epsilon^*) = 0$. Then

$$\begin{aligned} &g_k(A^*(\epsilon^*),\ B^*(\epsilon^*),\ \epsilon) \\ &\quad = g_k(A^*(\epsilon^*),\ B^*(\epsilon^*),\ \epsilon^*) - g_{2n+k}(B^*(\epsilon^*))(\epsilon - \epsilon^*) \\ &\quad < 0 \\ &g_{n+k}(A^*(\epsilon^*),\ B^*(\epsilon),\ \epsilon^*) \\ &\quad = g_{n+k}(A^*(\epsilon^*),\ B^*(\epsilon^*),\ \epsilon^*) - g_{2n+k}(B^*(\epsilon^*))(\epsilon - \epsilon^*) \\ &\quad < 0 \\ &-g_{2n+k}(B^*(\epsilon^*)) < -c \\ &\qquad\qquad\qquad\qquad (k = 1,2,\ldots,n) . \end{aligned}$$

Let $\lambda = \underset{k}{\mathrm{Max}} \left[g_k(A^*(\epsilon^*),\ B^*(\epsilon^*),\ \epsilon),\ g_{n+k}(A^*(\epsilon^*),\ B^*(\epsilon^*),\ \epsilon) \right]$.

For $\alpha \ge 1$ consider $(\alpha A^*(\epsilon^*),\ \alpha B^*(\epsilon^*),\ \alpha\lambda)$. Then

$$g_k(\alpha A^*(\epsilon^*),\ \alpha B^*(\epsilon^*),\ \epsilon) \le \alpha\lambda$$

$$g_{n+k}(\gamma A^*(\epsilon^*),\ \alpha B^*(\epsilon^*),\ \epsilon) \le \alpha\lambda \qquad (k = 1,2,\ldots,n)$$

$$-g_{2n+k}(\alpha B^*(\epsilon^*)) \le -c \ .$$

Therefore $\lambda^*(\epsilon) \le \alpha\lambda$. Since $\lambda < 0$ and $\alpha \ge 1$, $\lambda^*(\epsilon)$ can be made arbitrarily small. Therefore $\lambda^*(\epsilon)$ is $-\infty$.

Property 4

If $\epsilon < \epsilon^*$ then

$$\epsilon < \epsilon^* \le \epsilon + \frac{\lambda^*(\epsilon)}{c} \tag{9}$$

and the optimum solution associated with $\lambda^*(\epsilon)$ yields a rational approximation $R(\vec{z})$ such that

$$\max_{1 \le k \le n} |f(\vec{z}_k) - R(\vec{z}_k)| \le \epsilon + \frac{\lambda^*(\epsilon)}{c} \quad . \tag{10}$$

Proof: Let $(A^*(\epsilon), B^*(\epsilon), \lambda^*(\epsilon))$ be an optimum solution of program (8) for $\epsilon < \epsilon^*$. Then

$$\begin{aligned} g_k(A^*(\epsilon), B^*(\epsilon), \epsilon) &\le \lambda^*(\epsilon) \\ g_{n+k}(A^*(\epsilon), B^*(\epsilon), \epsilon) &\le \lambda^*(\epsilon) \qquad (k = 1,2,\dots,n) \\ - g_{2n+k}(B^*(\epsilon)) &\le - c \ . \end{aligned} \tag{11}$$

Let

$$P^*_\epsilon(\vec{z}) = \sum_{i=0}^{M} A^*_i(\epsilon) P_i(\vec{z})$$

$$Q^*_\epsilon(\vec{z}) = \sum_{j=0}^{N} B^*_j(\epsilon) Q_j(\vec{z})$$

$$R^*_\epsilon(\vec{z}) = \frac{P^*_\epsilon(\vec{z})}{Q^*_\epsilon(\vec{z})} \quad .$$

Then (11) yields

$$\begin{aligned} - P^*_\epsilon(\vec{z}_k) + y_k Q^*_\epsilon(\vec{z}_k) - (\epsilon + \frac{\lambda^*(\epsilon)}{Q^*_\epsilon(\vec{z}_k)})Q^*_\epsilon(\vec{z}_k) &\le 0 \\ P^*_\epsilon(\vec{z}_k) - y_k Q^*_\epsilon(\vec{z}_k) - (\epsilon + \frac{\lambda^*(\epsilon)}{Q^*_\epsilon(\vec{z}_k)})Q^*_\epsilon(\vec{z}_k) &\le 0 \\ - Q^*_\epsilon(\vec{z}_k) &\le -c \end{aligned}$$

$$(k = 1,2,\dots,n) \quad .$$

Therefore

$$-R^*_\epsilon(\vec{z}_k) + y_k \le \epsilon + \frac{\lambda^*(\epsilon)}{Q^*_\epsilon(\vec{z}_k)} ,$$

$$R^*_\epsilon(\vec{z}_k) - y_k \le \epsilon + \frac{\lambda^*(\epsilon)}{Q^*_\epsilon(\vec{z}_k)} ,$$

and

$$|f(\vec{z}_k) - R^*_\epsilon(\vec{z}_k)| \le \epsilon + \frac{\lambda^*(\epsilon)}{Q^*_\epsilon(\vec{z}_k)} \le \epsilon + \frac{\lambda^*(\epsilon)}{c} , \quad (k = 1,2,\ldots,n) .$$

Hence we have $\epsilon < \epsilon^* \le \epsilon + \frac{\lambda^*(\epsilon)}{c}$.

At present, a better lower bound for ϵ^* than ϵ has not been found when $\epsilon < \epsilon^*$. The above approach also fails to give upper and lower bounds for ϵ^* when $\epsilon > \epsilon^*$, since by property 3, $\lambda^*(\epsilon)$ is $-\infty$ in this case.

§ 3

As was observed in the previous section, if $\epsilon > \epsilon^*$ then program (8) has an unbounded optimum solution. This difficulty can be avoided by considering the dual program of program (8), since, by the Duality Theorem, the dual is infeasible when $\epsilon > \epsilon^*$. In this case an upper bound for ϵ^* as a function of ϵ can be obtained.

The dual program statement follows.

$$(12)\left\{\begin{array}{l} \text{Minimize } w\gamma \\ \text{subject to the constraints,} \\ (u,\ v,\ w) \left[\begin{bmatrix} -P & yQ & -1 \\ P & -yQ & -1 \\ 0 & -Q & 0 \end{bmatrix} + \epsilon \begin{bmatrix} 0 & -Q & 0 \\ 0 & -Q & 0 \\ 0 & 0 & 0 \end{bmatrix}\right] = (0,0,\ldots,0,-1) \\ \text{and} \\ u \ge 0 , \quad v \ge 0 , \quad w \ge 0 . \end{array}\right.$$

Here u, v, and w are unknown row vectors with n elements and

$$\gamma = \begin{bmatrix} -c \\ -c \\ \cdot \\ \cdot \\ \cdot \\ -c \end{bmatrix} \quad (\text{ n elements}) \; .$$

Normally, n is much larger than M + N + 3 and the dual program will be solved in preference to the primal problem.

We consider the following modification of program (12), which can be used to test the feasibility of (12).

$$(13)\begin{cases} \text{Minimize} \quad \pi = \sum_{i=0}^{M} \mu_i + \sum_{j=0}^{N} \nu_j + \tau \\ \text{subject to the constraints,} \\ (u,v,w,\mu,\nu,\tau)\left[\begin{bmatrix} P & -yQ & 1 \\ -P & yQ & 1 \\ 0 & Q & 0 \\ 1 & 0 \; . \; . \; . \; . & 0 \\ 0 & 1 \; . \; . \; . \; . & 0 \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & & \cdot \; \cdot \\ 0 & . \; . \; . \; . \; . & 1 \end{bmatrix} + \epsilon \begin{bmatrix} 0 & Q & 0 \\ 0 & Q & 0 \\ 0 & 0 & 0 \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ 0 & 0 & 0 \end{bmatrix}\right] = (0,0,\ldots,0,1) \\ \text{and} \\ u \geq 0 \;,\quad v \geq 0 \;,\quad w \geq 0 \;,\quad \mu \geq 0 \;,\quad \nu \geq 0 \;,\quad \tau \geq 0 \;, \end{cases}$$

where μ is $1 \times (M+1)$, ν is $1 \times (N+1)$, τ is 1×1 unknown row vector; respectively. The addition of the above identity matrix provides a starting basis and a means for detecting whether or not program (12) is feasible, and hence whether or not $\epsilon > \epsilon^*$. The following group of properties of π^* , the optimum value of π in (13), shows that there is a feasible solution of program (12) if and only if $\pi^* = 0$. Since $\pi \geq 0$, π^* is always finite.

Property 5

Program (13) is feasible.

Proof: $(u, v, w, \mu, \nu, \tau) = (0, 0, 0, 0, 0, 1)$ is a feasible solution of program (13).

Property 6

$\pi^*(\epsilon)$ is finite.

Proof: In the feasible solution given in Property 5, $\pi = 1$. Also, $\pi \geq 0$, since $\mu \geq 0$, $\nu \geq 0$, and $\tau \geq 0$. Therefore,

$$1 \geq \pi^*(\epsilon) \geq 0 \ .$$

Property 7

$\pi^*(\epsilon) = 0$ if and only if $\epsilon \leq \epsilon^*$. When $\epsilon > \epsilon^*$ and $0 < \pi^*(\epsilon) < 1$, then $\pi^*(\epsilon)$ is a strictly increasing function of ϵ .

Proof: If $\epsilon \leq \epsilon^*$, $\lambda^*(\epsilon) \geq 0$ and is finite. Therefore program (12) is feasible and hence there exists a feasible solution of program (13) such that $(\mu, \nu, \tau) = 0$ (i.e., $\pi^*(\epsilon)=0$) . If $\pi^*(\epsilon) = 0$, the above argument is reversible and hence by Property 3, $\epsilon \leq \epsilon^*$. Therefore $\pi^*(\epsilon) = 0$ if and only if $\epsilon \leq \epsilon^*$.

In order to prove the increasing property of $\pi^*(\epsilon)$, let $\epsilon > \epsilon^*$ and consider the following dual program of program (13).

$$\text{(14a)}\left\{\begin{array}{l}
\text{Maximize } \rho \\
\text{subject to the following constraints,} \\
\left[\begin{bmatrix}
P & -yQ & & & & 1 \\
-P & yQ & & & & 1 \\
0 & Q & & & & 0 \\
1 & 0 & \cdot & \cdot & \cdot & 0 \\
0 & 1 & \cdot & \cdot & \cdot & 0 \\
\cdot & \cdot & & & & \cdot \\
\cdot & & \cdot & & & \cdot \\
\cdot & & & \cdot & & \cdot \\
\cdot & & & & \cdot & \cdot \\
0 & \cdot & \cdot & \cdot & \cdot & 1
\end{bmatrix} + \epsilon \begin{bmatrix}
0 & Q & 0 \\
0 & Q & 0 \\
0 & 0 & 0 \\
\cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot \\
0 & 0 & 0 \\
\cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot \\
0 & 0 & 0
\end{bmatrix}\right] \begin{bmatrix} A \\ B \\ \rho \end{bmatrix} \leq \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ 1 \end{bmatrix} \left.\vphantom{\begin{matrix}1\\1\\ \cdot\\ \cdot\\ \cdot\\ \cdot\\1\end{matrix}}\right\} \text{(M+N+3) 1's}
\end{array}\right.$$

or

$$
(14b)\begin{cases}
\text{Maximize } \rho \\
\text{subject to the constraints,} \\
\qquad P(\vec{z}_k) - y_k Q(\vec{z}_k) + \epsilon Q(\vec{z}_k) + \rho \le 0 \\
\qquad -P(\vec{z}_k) + y_k Q(\vec{z}_k) + \epsilon Q(\vec{z}_k) + \rho \le 0 \\
\qquad Q(\vec{z}_k) \le 0 \\
\qquad A_i \le 1 \\
\qquad B_j \le 1 \\
\qquad \rho \le 1 \\
(i = 0,1,\dots,M \; ; \; j = 0,1,\dots,N \; ; \; k = 1,2,\dots,n) \; .
\end{cases}
$$

Let $\rho^*(\epsilon)$ denote the optimum value of ρ . By the Duality Theorem, $\rho^*(\epsilon) = \pi^*(\epsilon)$. We now show that $\rho^*(\epsilon)$ is a strictly increasing function of ϵ when $\epsilon > \epsilon^*$ and $0 < \rho^*(\epsilon) < 1$. First observe that $\rho^*(\epsilon) > 0$ if $\epsilon > \epsilon^*$. This follows from the fact that program (12) does not have a feasible solution if $\epsilon > \epsilon^*$; if program (12) is not feasible, $\pi^*(\epsilon) > 0$ for program (13); thus $\rho^*(\epsilon) > 0$ for program (14b). Next observe that $Q(\vec{z}_k) < 0$ for all k if $\rho^*(\epsilon) > 0$. For suppose that $\rho^*(\epsilon) > 0$ and $Q(\vec{z}_k) = 0$ for some k . Then $P(\vec{z}_k) + \rho \le 0$, $-P(\vec{z}_k) + \rho \le 0$. This implies that $\rho \le 0$, and therefore $\rho^*(\epsilon) \le 0$. This contradiction proves that $Q(\vec{z}_k) < 0$ for all k . Now let asterisks designate values corresponding to an optimum solution. Then for $\Delta\epsilon > 0$

$$P^*(\vec{z}_k) - y_k Q^*(\vec{z}_k) + (\epsilon + \Delta\epsilon) Q^*(\vec{z}_k) + \rho^*(\epsilon) - \Delta\epsilon Q^*(\vec{z}_k) \le 0$$

$$-P^*(\vec{z}_k) + y_k Q^*(\vec{z}_k) + (\epsilon + \Delta\epsilon) Q^*(\vec{z}_k) + \rho^*(\epsilon) - \Delta\epsilon Q^*(\vec{z}_k) \le 0 \; .$$

Therefore if we choose

$$\rho(\epsilon + \Delta\epsilon) = \text{Min}\left\lfloor 1, \underset{k}{\text{Min}}(\rho^*(\epsilon) - \Delta\epsilon Q^*(\vec{z}_k))\right\rceil ,$$

then $(A^*(\epsilon)$, $B^*(\epsilon)$, $\rho(\epsilon + \Delta\epsilon))$ satisfies program (14b) for $\epsilon + \Delta\epsilon$. Therefore $\rho^*(\epsilon + \Delta\epsilon) \geq \rho(\epsilon + \Delta\epsilon)$. Since $Q^*(\vec{z}_k) < 0$ for all k and $0 < \rho^*(\epsilon) < 1$, $\rho(\epsilon + \Delta\epsilon) > \rho^*(\epsilon)$. Hence $\rho^*(\epsilon + \Delta\epsilon) > \rho^*(\epsilon)$ and thus $\pi^*(\epsilon + \Delta\epsilon) > \pi^*(\epsilon)$ when $\epsilon > \epsilon^*$ and $0 < \rho^*(\epsilon) < 1$.

Property 8

When $\epsilon > \epsilon^*$, the optimum solution of program (13) yields a rational function $R(\vec{z}) = P(\vec{z})/Q(\vec{z})$ such that

$$(15) \qquad \begin{aligned} & Q(\vec{z}_k) < 0 \\ \text{and} \qquad & |f(\vec{z}_k) - R(\vec{z}_k)| \leq \epsilon + \frac{\pi^*(\epsilon)}{Q(\vec{z}_k)} \qquad (k = 1,2,\ldots,n) \end{aligned}$$

Hence

$$(16) \qquad \epsilon^* \leq \epsilon + \frac{\pi^*(\epsilon)}{\underset{k}{\text{Min}}\ Q(\vec{z}_k)} \quad .$$

Proof:

$$(17) \qquad \begin{aligned} P^*(\vec{z}_k) - y_k Q^*(\vec{z}_k) + \epsilon Q^*(\vec{z}_k) + \rho^*(\epsilon) \leq 0 \\ -P^*(\vec{z}_k) + y_k Q^*(\vec{z}_k) + \epsilon Q^*(\vec{z}_k) + \rho^*(\epsilon) \leq 0 \end{aligned}$$

holds, where

$$P^*(\vec{z}_k) = \sum_{i=0}^{M} A_i^* P_i(\vec{z}_k)$$

$$Q^*(\vec{z}_k) = \sum_{j=0}^{N} B_j^* Q_j(\vec{z}_k) \quad ,$$

and A^* and B^* are obtained from the inverse of the optimum basis of program (13). Remembering that $Q^*(\vec{z}_k) < 0$ when $\epsilon > \epsilon^*$, divide (17) by $Q^*(\vec{z}_k)$. Then

$$R^*(\vec{z}_k) - y_k + \epsilon + \frac{\rho^*(\epsilon)}{Q^*(\vec{z}_k)} \geq 0$$

$$-R^*(\vec{z}_k) + y_k + \epsilon + \frac{\rho^*(\epsilon)}{Q^*(\vec{z}_k)} \geq 0 \quad ,$$

where $R^*(\vec{z}) = P^*(\vec{z})/Q^*(\vec{z})$. This yields

$$\begin{aligned} |f(\vec{z}_k) - R^*(\vec{z}_k)| &\leq \epsilon + \frac{\rho^*(\epsilon)}{Q^*(\vec{z}_k)} = \epsilon + \frac{\pi^*(\epsilon)}{Q^*(\vec{z}_k)} \\ &\leq \epsilon + \frac{\pi^*(\epsilon)}{\underset{k}{\mathrm{Min}}\ Q^*(\vec{z}_k)} \end{aligned}$$

$$(k = 1,2,\ldots,n) \ .$$

§ 4

In §2 and §3, linear programs have been introduced whose solution enables upper bounds for ϵ^* to be obtained. Using these bounds, a method for iterating on ϵ is developed. Program (13), whose solution is always finite, is solved for a given value of ϵ . The corresponding value of $\pi^*(\epsilon)$ indicates whether $\epsilon > \epsilon^*$ or $\epsilon \leq \epsilon^*$, by property 7. If $\epsilon > \epsilon^*$, a better upper bound for ϵ^* than ϵ is given by property 8. If $\epsilon \leq \epsilon^*$, program (12) is solved and a value of $\lambda^*(\epsilon)$ is obtained by duality, and property 4 gives a bound for ϵ^* . The new value of ϵ is then chosen as the mid-point of the interval containing ϵ^*. When ϵ^* has been approached to within the desired precision, the optimum solution of the appropriate primal or dual program gives the coefficients of the approximating rational function. In general, it is impossible to obtain the best rational approximation exactly in a finite number of iterations.

Method I

(a) Set $i = 0$ and $\epsilon_0 = 0$.

(b) Solve program (13).

If $\pi^*(\epsilon_i) = 0$, go to (c) .

If $\pi^*(\epsilon_i) > 0$, set $\xi_i = \begin{cases} 0 & \text{for } i = 0 \\ \xi_{i-1} & \text{for } i \neq 0 \end{cases}$

$$\text{and } \eta_i = \epsilon_i + \frac{\pi^*(\epsilon_i)}{\operatorname*{Min}_k Q(\vec{z}_k)} .$$

Go to (d).

(c) Solve program (12)

If $\lambda^*(\epsilon_i) = 0$, halt. (The best rational approximation has been obtained.)

If $\lambda^*(\epsilon_i) > 0$, set $\xi_i = \epsilon_i$

$$\text{and } \eta_i = \begin{cases} \epsilon_i + \dfrac{\lambda^*(\epsilon_i)}{c} & \text{for } i = 0 \\ \operatorname{Min} \left[\eta_{i-1}, \ \epsilon_i + \dfrac{\lambda^*(\epsilon_i)}{c} \right] & \text{for } i \neq 0. \end{cases}$$

Go to (d).

(d) Set $\epsilon_{i+1} = \dfrac{\xi_i + \eta_i}{2}$, increase i by one and go to (b) .

The proof that the sequence $\{\epsilon_i\}_{i=0}^{\infty}$ defined by Method I converges to ϵ^* follows.

When $\pi^*(\epsilon_i) > 0$ and $i \neq 0$, we have $\epsilon_i > \epsilon^*$, and

$$\eta_i - \xi_i = \epsilon_i + \frac{\pi^*(\epsilon_i)}{\operatorname*{Min}_k Q(\vec{z}_k)} - \xi_{i-1} < \epsilon_i - \xi_{i-1}$$

$$= \frac{\xi_{i-1} + \eta_{i-1}}{2} - \xi_{i-1} = \frac{\eta_{i-1} - \xi_{i-1}}{2} .$$

It can be shown by induction that the interval $[\xi_i , \eta_i]$ contains ϵ^* by property 8.

When $\pi^*(\epsilon_i) = 0$, $\lambda^*(\epsilon_i) > 0$ and $i \neq 0$, we have $\epsilon_i < \epsilon^*$ and

$$\begin{aligned}\eta_i - \xi_i &= \text{Min}\left[\eta_{i-1} , \epsilon_i + \frac{\lambda^*(\epsilon_i)}{c}\right] - \epsilon_i \\ &\leq \eta_{i-1} - \epsilon_i \\ &= \eta_{i-1} - \frac{\eta_{i-1} + \xi_{i-1}}{2} = \frac{\eta_{i-1} - \xi_{i-1}}{2} .\end{aligned}$$

It can be shown by induction that the interval $[\xi_i , \eta_i]$ contains ϵ^* by properties 7 and 4. Therefore in both cases, if $\epsilon_i \neq \epsilon^*$, $\eta_i - \xi_i \leq \frac{(\eta_{i-1} - \xi_{i-1})}{2}$. Since the interval $[\xi_i, \eta_i]$ contains ϵ^* and $\epsilon_{i+1} = \frac{\eta_i + \xi_i}{2}$,

$$|\epsilon_{i+1} - \epsilon^*| \leq \frac{(\eta_i - \xi_i)}{2} \leq \frac{1}{2^{i+1}} (\eta_0 - \xi_0) = \frac{\lambda^*(0)}{2^{i+1}}$$

Hence $\{\epsilon_i\}_{i=0}^{\infty}$ converges to ϵ^* . If the first method given in Section IV of [1] is modified so that $\Delta\epsilon_{i+1} = \Delta\epsilon_i/2$ in all cases, then $|\epsilon^* - \epsilon_{i+1}| \leq \frac{\epsilon_0}{2^{i+1}}$. Hence, if $\frac{\lambda^*(0)}{c2^{i+1}}$ and $\frac{\epsilon_0}{2^{i+1}}$ are used as the stopping criteria for the respective methods, then the respective number of iterations required to obtain a given accuracy will depend on the relative values of $\lambda^*(0)/c$ and $\epsilon_0 = \underset{k}{\text{Max}}|f(\vec{z}_k)|$.

A second method for obtaining a convergent sequence $\{\epsilon_i\}_{i=0}^{\infty}$ is now given. A proof that $\{\epsilon_i\}_{i=0}^{\infty}$ converges to ϵ^* has not yet been obtained.

<u>Method II</u>

(a) Solve program (12) with $\epsilon = 0$, and set $\epsilon_0 = \frac{\lambda^*(0)}{c}$. Set i=0.

(b) Solve program (13) with $\epsilon = \epsilon_i$, and set

$$e_{i+1} = \epsilon_i + \frac{\pi^*(\epsilon_i)}{\underset{k}{\text{Min}}\ Q(\vec{z}_k)} .$$

(c) Increase i by one and go to (b) .

§ 5

In this section a more restricted program than program (7) will be considered. For this restricted program, both upper and lower bounds for ε^* can be obtained. Since both of these bounds are utilized in constructing the sequence of values of ε, the convergence to the optimum solution may be faster for this restricted program. The program is:

$$(18)\quad \begin{cases} \text{Minimize } e \\ \text{subject to the constraints,} \\ \qquad g_k(A,B,e) \le 0 \\ \qquad g_{n+k}(A,B,e) \le 0 \\ \qquad -g_{2n+k}(B) \le -c \qquad\qquad (k = 1,2,\ldots,n)\ , \\ \qquad g_{2n+k}(B) \le \bar{c} \end{cases}$$

where $\bar{c}$ is a real number greater than c. The fourth constraint of program (18) restricts the class of approximating functions. Suppose $c = \bar{c}$, i.e., $g_{2n+k}(B) = c$; then program (18) will give the best polynomial approximation. Therefore in the case of program (18), the class of approximating functions includes the set of all polynomials and is a subset of the set of all rational functions. It will be noted however that the fourth constraint might be added implicitly or explicitly in practice because of the finite length of computer words. If $\bar{c}$ is large enough so that the best rational approximation is included in the set of feasible solutions of program (18), then program (18) is the same as program (7). If e^* denotes the optimum e of program (18), then $\varepsilon^* \le e^*$.

The following program is considered, where e has a fixed positive value.

$$(19)\left\{\begin{array}{l}\text{Minimize } \lambda \\ \text{subject to the constraints,} \\ \qquad\qquad\qquad g_k(A,B,e) \le \lambda \\ \qquad\qquad\qquad g_{n+k}(A,B,e) \le \lambda \\ \qquad\qquad\qquad -g_{2n+k}(B) \le -c \qquad\qquad (k = 1,2,\dots,n) \\ \qquad\qquad\qquad g_{2n+k}(B) \le \bar{c} \end{array}\right.$$

Some useful properties of program (19) will now be stated.

Property 9

Program (19) is feasible if and only if there exists a vector B such that $c \le g_{2n+k}(B) \le \bar{c}$.

The proof of property 9 is similar to that of property 1.

Property 10

$\lambda^*(e)$ is a strictly decreasing function of e . $\lambda^*(e)$ is finite for every e .

Proof: The monotonicity of $\lambda^*(e)$ as a function of e can be proven in exactly the same manner as in the proof of property 2. The finiteness of $\lambda^*(e)$ is proven next. First observe that $\lambda^*(0) \ge 0$. For from program (19),

$$\begin{aligned} g_k(A,B,0) &\le \lambda^*(0) \\ g_{n+k}(A,B,0) &\le \lambda^*(0) \end{aligned} \qquad (k = 1,2,\dots,n) \ .$$

Since $g_k(A,B,0) = -g_{n+k}(A,B,0)$, by adding the two inequalities, $0 \le \lambda^*(0)$ is obtained. Also, either

$$\begin{aligned} g_k(A,B,e) &= g_k(A,B,0) - g_{2n+k}(B) \cdot e \\ &\ge -g_{2n+k}(B) \cdot e \qquad\qquad \text{or} \\ g_{n+k}(A,B,e) &= g_{n+k}(A,B,0) - g_{2n+k}(B) \cdot e \\ &\ge -g_{2n+k}(B) \cdot e \end{aligned}$$

for some k, since $0 \le \lambda^*(0)$.

Therefore

$$\underset{k}{\text{Max}}\left[g_k(A,B,e),\ g_{n+k}(A,B,e)\right] \geq -\ g_{2n+k}(B)\cdot e$$
$$\geq -\ \bar{c}\cdot e$$

Hence $$\lambda^*(e) \geq -\ \bar{c}\cdot e$$

Thus $\lambda^*(e)$ is finite.

Property 11

(i) $\lambda^*(e) > 0$ if and only if $e < e^*$.

(ii) $\lambda^*(e) = 0$ if and only if $e = e^*$.

(iii) $\lambda^*(e) < 0$ if and only if $e > e^*$.

Proof: (i) and (ii) can be proved in exactly the same manner as in the proof of property 3. Since $\lambda^*(e)$ is finite, (iii) follows directly from (i) and (ii).

Property 12

$$(20)\qquad \begin{aligned} e + \frac{\lambda^*(e)}{c} \leq e^* \leq e + \frac{\lambda^*(e)}{\bar{c}} \qquad & (\lambda^*(e) \leq 0) \\ e + \frac{\lambda^*(e)}{\bar{c}} \leq e^* \leq e + \frac{\lambda^*(e)}{c} \qquad & (\lambda^*(e) > 0). \end{aligned}$$

Proof: Suppose $e \geq e^*$. Then $\lambda^*(e) \leq 0$. Program (13) can be formulated as:

$$(21)\quad \begin{cases} \text{Maximize } \Delta e \\ \text{subject to the constraints,} \\ \qquad g_k(A,B,e - \Delta e) \leq 0 \\ \qquad g_{n+k}(A,B,e - \Delta e) \leq 0 \qquad (k = 1,2,\ldots,n) \\ \qquad c \leq g_{2n+k}(B) \leq \bar{c} \end{cases}$$

The first two constraints of program (21) will be rewritten as

$$\text{(22)} \qquad \begin{aligned} g_k(A,B,e) + g_{2n+k}(B) \cdot \Delta e &\le 0 \\ g_{n+k}(A,B,e) + g_{2n+k}(B) \cdot \Delta e &\le 0 \end{aligned} \qquad (k = 1,2,\ldots,n)$$

Since $g_{2n+k}(B) > c$ is required, (22) yields

$$\text{(23)} \qquad \begin{aligned} -\left[g_k(A,B,e)/g_{2n+k}(B)\right] &\ge \Delta e \\ -\left[g_{n+k}(A,B,e)/g_{2n+k}(B)\right] &\ge \Delta e \end{aligned} \qquad (k = 1,2,\ldots,n)$$

Let F be the set of all feasible solutions of program (18) (or(21)) and let Δe^* be the optimum Δe . Then

$$\Delta e^* = \sup_F \min_k \left[-\frac{g_k(A,B,e)}{g_{2n+k}(B)} \, , \; -\frac{g_{n+k}(A,B,e)}{g_{2n+k}(B)} \right]$$

$$= -\inf_F \max_k \left[\frac{g_k(A,B,e)}{g_{2n+k}(B)} \, , \; \frac{g_{n+k}(A,B,e)}{g_{2n+k}(B)} \right] .$$

Therefore

$$-\frac{\inf_F \max_k \left[g_k(A,B,e), g_{n+k}(A,B,e) \right]}{\bar{c}} \le \Delta e^* \le -\frac{\inf_F \max_k \left[g_k(A,B,e), g_{n+k}(A,B,e) \right]}{c} .$$

Hence

$$-\frac{\lambda^*(e)}{\bar{c}} \le \Delta e^* \le -\frac{\lambda^*(e)}{c} .$$

Since $e^* = e - \Delta e^*$,

$$e + \frac{\lambda^*(e)}{c} \le e^* \le e + \frac{\lambda^*(e)}{\bar{c}} \qquad (\lambda^*(e) \le 0) .$$

Next suppose $e \le e^*$. Then $\lambda^*(e) \ge 0$. Program (18) can be formulated as:

$$\text{(24)} \quad \left\{ \begin{array}{l} \text{Minimize } \Delta e' \\ \text{subject to the constraints,} \\ \qquad\qquad g_k(A,B,e + \Delta e') \le 0 \\ \qquad\qquad g_{n+k}(A,B,e + \Delta e') \le 0 \\ \qquad\qquad\qquad c \le g_{2n+k}(B) \le \bar{c} \end{array} \right.$$

By a similar argument to that given above, program (24) yields

$$\Delta e'^{*} = \underset{F}{\text{Inf}}\ \underset{k}{\text{Max}} \left[\frac{g_k(A,B,e)}{g_{2n+k}(B)} ,\ \frac{g_{n+k}(A,B,e)}{g_{2n+k}(B)} \right] ,$$

where $\Delta e'^{*}$ is the optimum $\Delta e'$.

Therefore

$$\frac{\underset{F}{\text{Inf}}\ \underset{k}{\text{Max}} \left[g_k(A,B,e),\ g_{n+k}(A,B,e) \right]}{\bar{c}} \le \Delta e'^{*} \le \frac{\underset{F}{\text{Inf}}\ \underset{k}{\text{Max}} \left[g_k(A,B,e),\ g_{n+k}(A,B,e) \right]}{c}$$

Hence
$$\frac{\lambda^*(e)}{\bar{c}} < \Delta e'^{*} \le \frac{\lambda^*(e)}{c} .$$

Since $e^* = e + \Delta e'^{*}$,

$$e + \frac{\lambda^*(e)}{\bar{c}} \le e^* \le e + \frac{\lambda^*(e)}{c} \qquad (\lambda^*(e) \ge 0) .$$

Property 13

The optimum solution associated with $\lambda^*(e)$ yields a rational approximation $R(\vec{z})$ such that

$$\underset{k}{\text{Max}} |f(\vec{z}_k) - R(\vec{z}_k)| \le e + \frac{\lambda^*(e)}{c} \text{ if } \lambda^*(e) \ge 0 ,$$

and

$$\underset{k}{\text{Max}} |f(\vec{z}_k) - R(\vec{z}_k)| \le e + \frac{\lambda^*(e)}{\bar{c}} \text{ if } \lambda^*(e) < 0 .$$

Proof: Let the asterisk designate the values of the optimum solution as before. Then by the same argument as in property 4,

$$|f(\vec{z}_k) - R^*(\vec{z}_k)| \le e + \frac{\lambda^*(e)}{Q^*(\vec{z}_k)} \qquad (k = 1,2,\ldots,n) .$$

If $e \leq e^*$, then $\lambda^*(e) \geq 0$ and

$$|f(\vec{z}_k) - R^*(\vec{z}_k)| \leq e + \frac{\lambda^*(e)}{c} .$$

If $e \geq e^*$, then $\lambda^*(e) \leq 0$ and

$$|f(\vec{z}_k) - R^*(\vec{z}_k)| \leq e + \frac{\lambda^*(e)}{\bar{c}} .$$

§ 6

Making use of the properties stated in §5 , algorithms for approaching e^* and the best approximating function can be constructed.

Method A

(a) Set $i = 0$ and $e_0 = 0$.

(b) Solve program (18) or the dual.

If $\lambda^*(e_i) = 0$, then terminate.

If $\lambda^*(e_i) > 0$, set

$$\xi_i = \begin{cases} \dfrac{\lambda^*(e_i)}{\bar{c}} , & i = 0 \\ \text{Max}\left\lfloor \xi_{i-1}, e_i + \dfrac{\lambda^*(e_i)}{\bar{c}} \right\rceil , & i \neq 0 \end{cases}$$

$$\eta_i = \begin{cases} \dfrac{\lambda^*(e_i)}{c} , & i = 0 \\ \text{Min}\left\lfloor \eta_{i-1}, e_i + \dfrac{\lambda^*(e_i)}{c} \right. , & i \neq 0 , \end{cases}$$

and go to (c) .

If $\lambda^*(e_i) < 0$, set

$$\xi_i = \text{Max}\left\lfloor \xi_{i-1}, e_i + \frac{\lambda^*(e_i)}{c} \right\rceil$$

$$\eta_i = \text{Min}\left\lfloor \eta_{i-1}, e_i + \frac{\lambda^*(e_i)}{\bar{c}} \right\rceil ,$$

and go to (c) .

(c) Set $e_{i+1} = \dfrac{\xi_i + \eta_i}{2}$, increase i by one, and go to (b) .

Method B

(a) Set $i = 0$ and $e_0 = 0$.

(b) Solve program (18) or the dual. Set

$$e_{i+1} = e_i + \frac{\lambda^*(e_i)}{\bar{c}}$$

(c) Increase i by one and go to (b) .

Method C

(a) Set $i = 0$ and $e_0 = 0$.

(b) Solve program (18) or the dual.

$$e_{i+1} = \begin{cases} \dfrac{\lambda^*(e_i)}{c} , & i = 0 \\ e_i + \dfrac{\lambda^*(e_i)}{\bar{c}} , & i \neq 0 . \end{cases}$$

(c) Increase i by one and go to (b) .

Property 14

For Method A, $\lim\limits_{i \to \infty} e_i = e^*$ and $|e_{i+1} - e^*| \leq \dfrac{(\eta_0 - \xi_0)}{2^{i+1}}$, $i \neq 0$.

Proof: Since the interval $[\xi_i, \eta_i]$ contains e^* by property 12,

$$|e_{i+1} - e^*| \leq \frac{\eta_i - \xi_i}{2} .$$

If $\lambda^*(e) \geq 0$, $\xi_i \geq e_i + \dfrac{\lambda^*(e_i)}{\bar{c}}$ and $\eta_i \leq \eta_{i-1}$.

Hence $$\eta_i - \xi_i \leq \eta_{i-1} - \left[e_i + \frac{\lambda^*(e_i)}{\bar{c}}\right] \leq \eta_{i-1} - e_i$$

$$= \eta_{i-1} - \left[\frac{\xi_{i-1} + \eta_{i-1}}{2}\right] = \frac{\eta_{i-1} - \xi_{i-1}}{2} .$$

If $\lambda^*(e) < 0$, $\eta_i \le e_i + \frac{\lambda^*(e_i)}{\bar{c}}$ and $\xi_i \ge \xi_{i-1}$.

Hence
$$\eta_i - \xi_i \le e_i + \frac{\lambda^*(e_i)}{\bar{c}} - \xi_{i-1} \le e_i - \xi_{i-1}$$
$$= \frac{\xi_{i-1} + \eta_{i-1}}{2} - \xi_{i-1} = \frac{\eta_{i-1} - \xi_{i-1}}{2} .$$

Therefore in both cases, $|e_{i+1} - e^*| \le \frac{(\eta_0 - \xi_0)}{2^{i+1}}$, and hence $\lim_{i\to\infty} e_i = e^*$.

Property 15

For Method B, $\lim_{i\to\infty} e_i = e^*$, and $e^* - e_{i+1} \le (1/c - 1/\bar{c})\lambda^*(e_i)$.

Proof: $e_0 = 0 \le e^*$ and $\lambda^*(e_0) \ge 0$. Hence $e_1 \ge e_0$. By property 12, $e_i \le e^*$. It can be shown by induction on i , using properties 11 and 12, that $\lambda^*(e_i) \ge 0$ and $0 \le e_i \le e_{i+1} \le e^*$ for all i . Hence $\{e_i\}_{i=0}^{\infty}$ has a sequential limit. Thus $\lim_{i\to\infty} (e_{i+1}-e_i) = 0$ and $\lim_{i\to\infty} \lambda^*(e_i) = 0$, by construction of e_{i+1} . Since $\lambda^*(e)$ is a strictly decreasing function of e , $\lim_{i\to\infty} \lambda^*(e_i) = 0$ and $\lambda^*(e^*) = 0$, we have $\lim_{i\to\infty} e_i = e^*$. Also, $e_i + \frac{\lambda^*(e_i)}{\bar{c}} = e_{i+1} \le e^* \le e_i + \frac{\lambda^*(e_i)}{c}$ by property 12. Hence $e^* - e_{i+1} \le (1/c - 1/\bar{c})\lambda^*(e_i)$. Note that in Method B $\{e_i\}_{i=0}^{\infty}$ converges to e^* from below.

Property 16

For Method C, $\lim_{i\to\infty} e_i = e^*$ and $e_{i+1} - e^* \le -(1/c - 1/\bar{c})\lambda^*(e_i)$.

Proof: $e_0 = 0 \le e^*$ and $\lambda^*(e_0) \ge 0$. $e_1 = \lambda^*(e_0)/c \ge e^*$.
By an argument similar to that for property 15, $0 \le e^* \le e_{i+1} \le e_i$ for $i = 1,2,\ldots$. Hence, $\lim_{i\to\infty} e_i = e^*$.

Also $$e_i + \frac{\lambda^*(e_i)}{c} \leq e^* \leq e_{i+1} = e_i + \frac{\lambda^*(e_i)}{\bar{c}} \text{ , and}$$

hence $$e_{i+1} - e^* \leq (1/\bar{c} - 1/c)\lambda^*(e_i) \text{ .}$$

Note that in Method C, $\{e_i\}_{i=0}^{\infty}$ converges to e^* from above.

REFERENCE

1. Suzuki, S., "The Application of Linear Programming Techniques to Rational Approximation Problems," Progress Report No. 4, MSFC Report No. MTP-AERO-63-65, September 19, 1963.

COMPUTATION CENTER
UNIVERSITY OF NORTH CAROLINA

INVERSE ESTIMATION

by

G. W. Adkins

CHAPEL HILL, NORTH CAROLINA

COMPUTATION CENTER
UNIVERSITY OF NORTH CAROLINA
CHAPEL HILL, NORTH CAROLINA

INVERSE ESTIMATION

By

G. W. Adkins

SUMMARY

20960 A

This paper briefly describes the procedures employed in the application of inverse estimation to problems in experimental design, shows how this technique might be applied in the development of guidance function approximations, and indicates the problem areas that must be investigated for such an application.

Author

INVERSE ESTIMATION

1. Introduction

When an estimate of a (vector) function such as

$$\underline{x} = g(\underline{y}) \tag{1.1}$$

is desired, the usual procedure is to select a set of sample values for $\underline{y}$, observe the responses $\underline{x}$ and use this data to estimate the desired function. However, there are times when the $\underline{x}$'s , and not the $\underline{y}$'s , are at our disposal as independent variables. In this case, it may be more convenient to fit the function

$$\underline{y} = f(\underline{x}) \tag{1.2}$$

according to some criterion of "best fit" then, using this relationship, determine the "best" estimate of the inverse relationship (1.1). In this paper, the problem of inverse estimation is considered and some methods presented.

2. Direct Estimation

The usual problem in response surface estimation is determining the parameters in the model

$$\underline{\eta} = f(\underline{x}) \tag{2.1}$$

where $\underline{\eta}$ is a $(q \times 1)$ vector of dependent variables, or response variables, and $\underline{x}$ is a $(p \times 1)$ vector of controllable independent variables. If each element $f_i(\underline{x})$ of the vector $f(\underline{x})$ is a polynomial form,

$$
\begin{aligned}
f_i(\underline{x}) = {} & b_{i0}x_0^r + b_{i1}x_0^{r-1}x_1 + b_{i2}x_0^{r-1}x_2 \\
& + \ldots + b_{ip}x_0^{r-1}x_p + b_{i12}x_0^{r-2}x_1x_2 \\
& + \ldots + b_{i,p-1,p}x_0^{r-2}x_{p-1}x_p + \ldots + b_{i11}x_0^{r-2}x_1^2 \\
& + \ldots + b_{ipp}x_0^{r-2}x_p^2 + \ldots + b_{ipp\ldots p}x_p^r \ ,
\end{aligned}
\tag{2.2}
$$

and $x_0 \equiv 1$,

We can express each such element as

$$f_i(\underline{x}) = \underline{\beta}_i' \, \underline{x}^{[r]} \ , \tag{2.3}$$

where the vector $\underline{x}^{[r]}$ contains elements of the general form

$$x_0^{m_0} x_1^{m_1} \ldots x_p^{m_p} \tag{2.4}$$

and

$$m_0 + m_1 + \ldots + m_p = r. \tag{2.5}$$

Then model (2.1) can be stated as

$$\underline{\eta} = B' \, \underline{x}^{[r]} \tag{2.6}$$

and is amenable to the experimental design methods of Box and Hunter [1] and Bose and Draper [2].

Having chosen a design, i.e., observation points $\underline{x}_j$, $j = 1,2,\ldots,n,$ and observed the responses,

$$\underline{y}_j = \underline{\eta}_j + \underline{\varepsilon}_j = f(\underline{x}_j) + \underline{\varepsilon}_j \ , \tag{2.7}$$

where $\underline{\varepsilon}_j$ is an error vector, we may write the design model as

$$
\begin{aligned}
\underline{y}_1' &= \underline{x}_1^{[r]'} B + \underline{\varepsilon}'_1 \\
\underline{y}_2' &= \underline{x}_2^{[r]'} B + \varepsilon'_2 \\
&\ \ \vdots \\
\underline{y}_n' &= \underline{x}_n^{[r]'} B + \underline{\varepsilon}_n'
\end{aligned}
\tag{2.8}
$$

or

$$Y = XB + E \tag{2.9}$$

where Y is a $(n \times q)$ matrix of response variables, X is the $(n \times m)$ design matrix (m determined by r), B is the $(m \times q)$ matrix of unknown coefficients, and E is the $(n \times q)$ matrix of error terms. Using the least-squares principle, B is estimated by $\hat{B}$, i.e., that matrix which minimizes the trace of $E'E$. It can be shown that

$$\hat{B} = [X'X]^{-1}X'Y \; . \tag{2.10}$$

In statistical terms, if the error vectors $\underline{\varepsilon}_i$ are assumed, as usual, to follow the multivariate normal frequency distribution,

$$f(\underline{e}) = \frac{|V|^{-1}}{(\pi)^{n/2}} \exp\{- 1/2[\underline{e}'V^{-1}\underline{e}]\} \; , \tag{2.11}$$

and furthermore the $\underline{\varepsilon}_i$ are assumed to be stochastically independent, then the estimate $\hat{B}$ is the maximum liklihood, minimum variance estimate. The covariance matrix V is estimated by,

$$\hat{V} = \frac{1}{n-p-1} \{(Y - XB)'(Y - XB)\} \quad . \tag{2.12}$$

An element v_{ik} of $\hat{V}$ measures the simultaneous "lack of fit" of y_i and y_k .

For any point $\underline{x}$, not necessarily a design point, the response vector $\underline{\eta}$ is estimated by

$$\underline{\hat{\eta}} = \underline{\hat{y}} = \hat{B}'\underline{x}^{[r]} \tag{2.13}$$

and the covariance matrix for this estimate is

$$\left\{\underline{x}^{[r]'}[X'X]^{-1}x^{[r]}\right\} V \tag{2.14}$$

3. The Problem of Inverse Estimation

There are times when we want to estimate not the function expressed in (2.1), but the inverse function,

$$\underline{x} = g(\underline{\eta}) \quad , \tag{3.1}$$

where, again, the $\underline{x}$'s are the controllable variables and the $\underline{\eta}$'s are the response variables. For example, consider the problem of calibrating an hourglass. Let t_0 be the elapsed time measured by the hourglass and t be the true elapsed time as measured by some presumably unerring device. An estimate of the function

$$t = g(t_0) \tag{3.2}$$

is desired. The logical way to approach this problem would be to choose times $t_1, t_2, \ldots, t_n$ and at each time, t_i , observe the hourglass time $(t_{0i} + \varepsilon_i)$ and then use this data to estimate $g(t_0)$. This is an inverse estimation problem. Our data reflects a model of type (2.1) where t_0 is the response variable and t is the chosen independent variable, yet we want to estimate a function of type (3.2).

Williams [3] [4] discusses the problem of inverse estimation in the simple linear case and considers the problem of estimating the quantities of two sugars, glucose and galactose, in a solution by observing the optical density of the solution to light of two different wavelengths. An estimating function is determined by using solutions of known sugar content and observing the associated optical densities, the response variables thus being the optical densities and the controllable variables the quantitative sugar contents.

In the flat earth problem discussed in [5] and the various progress reports [6], the problem is to determine an equation for estimating the optimum trajectory to continue, knowing the instantaneous conditions of position, velocity, propellant flowrate, and (F/m). The equation is to be determined from observations of the instantaneous conditions along chosen trajectories known to be optimum. Characterizing a trajectory by the steering function $\chi(c_0, c_1, c_2)$ and the cut-off time t_c, a model for the observed data is

$$\underline{w} = f(\underline{u}) \tag{3.3}$$

where $\underline{w}$ is the vector of observed conditions $(x + \epsilon_x, y + \epsilon_j, \dot{x} + \epsilon_{\dot{x}}, \dot{y} + \epsilon_{\dot{y}}, \mu + \epsilon_\mu, F/m + \epsilon_{F/m})$, the ϵ's perhaps in-flight measurement errors, and $\underline{u}$ is the vector of controllable variables (c_0, c_1, c_2, and t_c, the time of cut-off). Since the steering angle χ is a direct function [5] of c_0, c_1, and c_2, it may be feasible to include χ in the above model as a concomittant variable, creating a model analagous to the analysis of covariance model familiar to experiment analysts.

From the model (3.3), we want to estimate a working relationship

$$\underline{u} = g(\underline{w}) \tag{3.4}$$

That is, given measurements of the instantaneous conditions, we want to determine which optimum trajectory we are on and adjust the steering and cut-off mechanisms accordingly. One proposed method for doing this is to generate bundles of optimum trajectories, at points along these trajectories obtain values for $\underline{w}$, and then fit the model (3.4) by, say, a least squares procedure. Here, $\underline{w}$ becomes the vector of independent variables and $\underline{u}$ the vector of response variables when, in fact, their true roles are just the opposite.

Since the vector $\underline{w}$ is not controllable, the choice of trajectories and observation points on these trajectories becomes a matter of personal judgment rather than a matter of mathematical choice. In the direct estimation procedure discussed in the previous section one generally chooses design points such that the matrix in (2.10) is easily inverted, the estimates in (2.10) are relatively free from interaction and correlation with one another, and the covariance matrix in (2.14) satisfies some specified criterion. However, if the variables considered independent are not actually controllable, we are not free to choose points that satisfy these criteria.

Therefore, the question arises as to whether or not it may be better to fit the model (3.3), where the vector $\underline{u}$ is controllable, by direct estimation and then solve (3.4) by the inverse estimation techniques of the next section.

4. Determining the Estimates

Given the linear model

$$\underline{y} = \underline{\eta} + \underline{\varepsilon} = B'\underline{x} + \underline{\varepsilon} \tag{4.1}$$

we can determine the estimates $\hat{B}$ and $\hat{V}$ from observed data. The estimated liklihood, then, is

$$\frac{|\hat{V}|^{-1}}{(2\pi)^{n/2}} \exp\{-1/2[\underline{e}'\hat{V}^{-1}\underline{e}]\} \tag{4.2}$$

where $\underline{e} = \underline{y} - \hat{B}'\underline{x}$. If we wish to estimate $\underline{x}$ from an observation of $\underline{y}$, Williams [3] suggests that a reasonable criterion would be to choose the $\underline{x}$ that has minimum estimated variance, or that maximizes the estimated likelihood. From (4.2), we want to choose the $\underline{x}$ that minimizes the quadratic form

$$Q = (\underline{y}' - \underline{x}'\hat{B})\hat{V}^{-1}(\underline{y} - \hat{B}'\underline{x}) \ . \tag{4.3}$$

Differentiating with respect to $\underline{x}$, we get,

$$\hat{B}\hat{V}^{-1}\hat{B}'\hat{\underline{x}} - \hat{B}\hat{V}^{-1}\underline{y} = 0 \quad . \tag{4.4}$$

Henceforth, we shall disregard the carats and it will be understood that B, V, and $\underline{x}$ refer to the estimated quantities.

Solving equation (4.4) we find that

$$\underline{x} = [BV^{-1}B']^{*}BV^{-1}\underline{y} \quad , \tag{4.5}$$

where the conditional inverse $[BV^{-1}B']^{*}$ is defined by

$$[BV^{-1}B'][BV^{-1}B']^{*}[BV^{-1}B'] = [BV^{-1}B'] \quad . \tag{4.6}$$

If $[BV^{-1}B]$ is non-singular, then $\underline{x}$ is uniquely determined and

$$[BV^{-1}B']^{*} = [BV^{-1}B']^{-1} \quad . \tag{4.7}$$

Note that if $\underline{y}$ is $(p \times 1)$ and $\underline{x}$ is $(p \times 1)$, then B is a square matrix, and, further, if B is non-singular, (4.5) reduces to

$$\underline{x} = B'^{-1}VB^{-1}BV^{-1}\underline{y} = B'^{-1}\underline{y} \tag{4.8}$$

Writing (4.5) as

$$\underline{x} = G\underline{y} \tag{4.9}$$

we see the estimated variance of $\underline{x}$ is

$$V(\underline{x}) = GV(\underline{y})G' \tag{4.10}$$

and $V(\underline{y})$ can be determined from (2.14).

If the model is extended to include terms of higher order, i.e.,

$$y_i = \sum_j b_{ij}x_j + \sum_j\sum_k b_{ijk}x_jx_k + \sum_j\sum_k\sum_l b_{ijkl}x_jx_kx_l + \ldots + e_i$$

$$\text{or} \qquad \underline{y} = B'\,\underline{x}^{[r]} + \underline{e} \tag{4.11}$$

then the minimum of the quadratic form is not as explicitly determined since the equations $\partial Q/\partial\underline{x}$ will include non-linear terms in the x's . There is, however, a relatively simple iterative scheme to determine the minimum.

We can linearize (4.11) by redefining the variables as

$$\begin{aligned} z_1 &= x_1 \\ z_2 &= x_2 \\ &\vdots \\ z_p &= x_p \\ z_{p+1} &= x_1x_2 \\ z_{p+2} &= x_1x_3 \\ &\vdots \\ z_m &= x_p^r \quad . \end{aligned} \tag{4.12}$$

(4.11) becomes,

$$\underline{y} = B'\underline{z} + \underline{e} \,, \tag{4.13}$$

and the problem now is to minimize,

$$Q = (\underline{y}' - \underline{z}'B)V^{-1}(\underline{y} - B'\underline{z}) \,, \tag{4.14}$$

subject to the constraints,

$$\begin{aligned} z_{p+1} &= z_1z_2 \\ z_{p+2} &= z_1z_3 \\ &\vdots \\ z_m &= z_p^r \quad . \end{aligned} \tag{4.15}$$

Using the method of Lagrange multipliers, we derive the set of equations,

$$\begin{bmatrix} A_{p\times p} & C_{p\times s} & 0_{p\times s} \\ C'_{s\times p} & D_{s\times s} & I_{s\times s} \\ 0_{s\times p} & I_{s\times s} & 0_{s\times s} \end{bmatrix} \begin{bmatrix} \underline{z}_p \\ \underline{z}_s \\ \underline{\lambda} \end{bmatrix} = \begin{bmatrix} u_1(\underline{\lambda},\ z_p,\ \underline{y}) \\ u_2(\underline{y}) \\ u_3(\underline{z}_p) \end{bmatrix} \tag{4.16}$$

where the matrix $BV^{-1}B'$ has been partitioned into $\begin{bmatrix} A & C \\ C' & D \end{bmatrix}$, 0 is a matrix of zeros, I is the identity matrix, $\underline{z}_p$ is the vector of primary variables $x_1, x_2, \ldots, x_p$, $\underline{z}_s$ is the vector of secondary variables, i.e., the higher-order terms, $\underline{\lambda}$ is the vector of Lagrange multipliers, $\underline{u}_1$ is bilinear in $\underline{\lambda}$ and $\underline{z}_p$ but linear in $\underline{y}$, $\underline{u}_2$ is linear in $\underline{y}$, and $\underline{u}_3$ is the right-hand side of (4.15). Equations (4.16) can be solved iteratively by the following procedure:

$$\begin{bmatrix} \underline{z}_{p,k} \\ \underline{z}_{s,k} \\ \underline{\lambda}_k \end{bmatrix} = \begin{bmatrix} A^{-1}_{p\times p} & 0_{p\times s} & -A^{-1}C_{p\times s} \\ 0_{s\times p} & 0_{s\times s} & I_{s\times s} \\ -C'A'^{-1}_{s\times p} & I_{s\times s} & C'A^{-1}C-D \end{bmatrix} \begin{bmatrix} \underline{u}_{1,k-1} \\ \underline{u}_{2,k-1} \\ \underline{u}_{3,k-1} \end{bmatrix}$$

$$\begin{bmatrix} \underline{u}_{1,k} \\ \underline{u}_{2,k} \\ \underline{u}_{3,k} \end{bmatrix} = \begin{bmatrix} u_1(\underline{\lambda}_{k-1}, \underline{z}_{p,k-1}, \underline{y} \\ u_{2,k-1}(\underline{y}) \\ u_3(\underline{z}_{p,k-1}) \end{bmatrix} \tag{4.17}$$

Since $\underline{u}_{2,k}$ is invariant with respect to k and $\underline{z}_{s,k}$ is identically $\underline{u}_3$, (4.17) can be reduced to the set of m equations

$$\begin{aligned} \underline{z}_{p,k} &= A^{-1}u_1(\underline{\lambda}_{k-1}, \underline{z}_{p,k-1}, \underline{y}) - A^{-1}Cu_3(\underline{z}_{p,k-1}) \\ \underline{\lambda}_k &= -C'A'^{-1}u_1(\underline{\lambda}_{k-1}, \underline{z}_{p,k-1}, \underline{y}) + u_2(\underline{y}) + [C'A^{-1}C - D]u_3(\underline{z}_{p,k-1}) \end{aligned} \tag{4.18}$$

5. Discussion

If the inverse estimation procedure were used in the flat earth problem, we would first determine the relationship (3.3) using experimental design techniques with, say, a Runge-Kutta method to generate the response variables from the controllable variables c_0, c_1, c_2, and t_c. Then, in applying our model, we would measure the response variables x, y, $\dot{x}$, $\dot{y}$, u, (F/m) in flight and, using the procedure of (4.18), estimate the values of c_0, c_1, c_2, and t_c that would have given us this observed response.

On the surface this approach seems to be a logical attack on the problem. However, there are several critical questions concerning equations (4.18) that must be explored. The most obvious ones concern computing time, computer storage requirements, and the variance of the estimates obtained. At this point these questions have not been explored in detail but a first glance at the equations involved yields a first approximation answer of too long, too much, and too great for in-flight calculations by an on-board computer. Therefore, the difficulties in applying the results of this inverse estimation procedure might nullify the simplifications obtained in the direct estimation problem by the application of controllable experimental design.

REFERENCES

1. Box, G. E. P. and J. S. Hunter, "Multi-factor Experimental Designs for Exploring Response Surfaces," Annals of Math. Stat., Vol. 28, (1957), pp. 195-241.

2. Bose, R. C. and N. R. Draper, "Second Order Rotatable Designs in Three Dimensions," Annals of Math. Stat., Vol 30, No. 4 (1959) pp. 1097-1112.

3. Williams, E. J., "Simultaneous Regression Equation in Experimentation," Biometrika, Vol. 45 (1958) pp. 96-110.

4. Williams, E. J., Regression Analysis, J. Wiley & Sons, N. Y., (1959).

5. Hoelker, R. F. and W. E. Miner, "Introduction Into the Concept of Adaptive Guidance Modes," Aeroballistics Internal Note No. 21-60, MSFC.

6. Progress Reports 1-4 On Studies in the Fields of Space Flight and Guidance Theory, MSFC, Huntsville, Ala.

DEPARTMENT OF MATHEMATICS
NORTHEAST LOUISIANA STATE COLLEGE
MONROE, LOUISIANA

The Selection of a Least Squares Approximating Function to Satisfy a Given Error Tolerance

by

Daniel E. Dupree, F. L. Harmon, R. A. Hickman,
Edward B. Anders, James O'Neil

DEPARTMENT OF MATHEMATICS
NORTHEAST LOUISIANA STATE COLLEGE
Monroe, Louisiana

The Selection of a Least Squares Approximating Function to Satisfy a Given Error Tolerance

by

Daniel E. Dupree, F. L. Harmon, R. A. Hickman,
Edward B. Anders, James O'Neil

SUMMARY

20961

A

A technique for obtaining a function which yields an error, in the sense of least squares, that is less than a specified tolerance is developed.

Author

I. INTRODUCTION

In [1] , a recursion process was developed for obtaining the coefficients A_0, A_1, ... , A_N of the function $A_0\varphi_0(\beta) + A_1\varphi_1(\beta) + - - - + A_N\varphi_N(\beta)$ such that

$$E = \sum_{i=0}^{n} \left\{ X(\beta_i) - \sum_{j=0}^{N} A_j\varphi_j(\beta_i) \right\}^2$$

is minimum. This scheme yields the coefficients of the approximating function without having to solve the normal equations. Of course, the least squares procedure minimizes the sum of the squared errors, yet we have no assurance of the relative size of this error. In this paper, we will develop a process for choosing the approximating function in such a fashion that the error will not exceed a given tolerance.

Before doing this, let us examine more closely the error E incurred by using the function $\sum_{j=0}^{N} A_j\varphi_j(\beta)$ as an approximating function. If the vectors $\bar{\varphi}_0, \bar{\varphi}_1, \dots, \bar{\varphi}_N$, $N < n$, are used to obtain the collection $\bar{e}_0, \bar{e}_1, \dots, \bar{e}_N$ of orthonormal vectors as in [1], then the error E can be written as follows:

$$E = \sum_{i=0}^{n} \left[X(\beta_i) - \sum_{j=0}^{N} A_j\varphi_j(\beta_i) \right]^2 = \| \bar{X} - A_0\bar{\varphi}_0 - A_1\bar{\varphi}_1 - \dots - A_N\bar{\varphi}_N \|^2 = \| \bar{X} - \sum_{j=0}^{N} (\bar{X}, \bar{e}_j)\,\bar{e}_j \|^2 =$$

$$\| \bar{X} \|^2 - \left[\bar{X}, \sum_{j=0}^{N} (\bar{X}, \bar{e}_j)\,\bar{e}_j \right] - \left[\bar{X}, \sum_{j=0}^{N} (\bar{X}, \bar{e}_j)\,\bar{e}_j \right] + \| \sum_{j=0}^{N} (\bar{X}, \bar{e}_j)\,\bar{e}_j \|^2 = \| \bar{X} \|^2 - \sum_{j=0}^{N} (\bar{X}, \bar{e}_j)^2 .$$

From this representation of E, we are able to observe the following:

1) $\| \bar{X} \|^2$ is an upper bound for E.

2) A sum of any k of the N + 1 terms $(\bar{X}, \bar{e}_j)^2$, $0 < k < N + 1$, will yield an error $E' > E$.

3) If $\bar{e}_{N+1}$ is any other non-zero vector orthogonal to each of $\bar{e}_0, \bar{e}_1, \dots, \bar{e}_N$, then $\| \bar{X} \|^2 - \sum_{j=0}^{N+1} (\bar{X}, \bar{e}_j)^2 < E$.

II. SELECTION OF THE FUNCTION

After evaluating $\| \overline{X} \|^2 - \sum_{j=0}^{N} (\overline{X}, \overline{e}_j)^2$, we may find that this value still exceeds a given error tolerance δ. Then we wish to find $\overline{\varphi}_{N+1}$ such that

$$\| \overline{X} \|^2 - \sum_{j=0}^{N} (\overline{X}, \overline{e}_j)^2 - (\overline{X}, \overline{e}_{n+1})^2 \leq \delta;$$

i.e., find $\overline{\varphi}_{N+1}$ such that

$$(\overline{X}, \overline{e}_{n+1})^2 \geq \| \overline{X} \|^2 - \sum_{j=0}^{N} (\overline{X}, \overline{e}_j)^2 - \delta,$$

where $\overline{e}_{n+1}$ is the vector associated with $\overline{\varphi}_{N+1}$ that is orthogonal to $\overline{e}_0, \overline{e}_1, \dots, \overline{e}_N$.

Suppose we let

$$\overline{\varphi}_{N+1} = (\lambda_0, \lambda_1, \dots, \lambda_n).$$

Then

$$\begin{aligned} \overline{\varphi}'_{N+1} &= \overline{\varphi}_{N+1} - (\overline{\varphi}_{N+1}, \overline{e}_0)\, \overline{e}_0 - \dots - (\overline{\varphi}_{N+1}, \overline{e}_N)\, \overline{e}_N \\ &= (\lambda_0, \lambda_1, \dots, \lambda_n) - \Big(\sum_{i=0}^{n} \lambda_i\, e_{0i} \Big) \overline{e}_0 - \dots \\ &\quad - \Big(\sum_{i=0}^{n} \lambda_i\, e_{Ni} \Big) \overline{e}_N, \end{aligned}$$

if $\overline{e}_j = (e_{j0}, e_{j1}, \dots, e_{jn})$, $j = 0, 1, \dots, N$.

Therefore,

$$\bar{e}_{N+1} = \frac{(\lambda_0,\lambda_1,\ldots,\lambda_n) - (\sum_{i=0}^{n} \lambda_i\, e_{0i})\,\bar{e}_0 - \ldots - (\sum_{i=0}^{n} \lambda_i\, e_{Ni})\,\bar{e}_N}{\left\{\sum_{i=0}^{n} \lambda_i^2 - (\sum_{i=0}^{n} \lambda_i\, e_{0i})^2 - \ldots - (\sum_{i=0}^{n} \lambda_i\, e_{Ni})^2\right\}^{1/2}} ,$$

and if $\bar{X} = (t_0,t_1,\ldots,t_n)$, then

$$(\bar{X},\, \bar{e}_{N+1})^2 = \frac{\left[\sum_{i=0}^{n} \lambda_i\, t_i - (\sum_{i=0}^{n} \lambda_i\, e_{0i})(\bar{X},\, \bar{e}_0) - \ldots - (\sum_{i=0}^{n} \lambda_i\, e_{Ni})(\bar{X},\, \bar{e}_N)\right]^2}{\sum_{i=0}^{n} \lambda_i^2 - (\sum_{i=0}^{n} \lambda_i\, e_{0i})^2 - \ldots - (\sum_{i=0}^{n} \lambda_i\, e_{Ni})^2} .$$

Thus, to have

$$(\bar{X},\, \bar{e}_{N+1})^2 \geq \| \bar{X} \|^2 - \sum_{j=0}^{N} (\bar{X},\, \bar{e}_j)^2 - \delta,$$

we must have

$$\left[\sum_{i=0}^{n} \lambda_i\, t_i - (\sum_{i=0}^{n} \lambda_i\, e_{0i})(\bar{X},\, \bar{e}_0) - \ldots - (\sum_{i=0}^{n} \lambda_i\, e_{Ni})(\bar{X},\, \bar{e}_N)\right]^2 \geq$$

$$\left[\sum_{i=0}^{n} \lambda_i^2 - (\sum_{i=0}^{n} \lambda_i\, e_{0i})^2 - \ldots - (\sum_{i=0}^{n} \lambda_i\, e_{Ni})^2\right]\left[\| \bar{X} \|^2 - \sum_{j=0}^{N} (\bar{X},\, \bar{e}_j)^2 - \delta\right],$$

or

$$\left[\sum_{i=0}^{n} \lambda_i\, \{t_i - (\bar{X},\, \bar{e}_0)\, e_{0i} - \ldots - (\bar{X},\, \bar{e}_N)\, e_{Ni}\}\right]^2 \geq$$

$$\left[\sum_{i=0}^{n} \lambda_i^2 - (\sum_{i=0}^{n} \lambda_i\, e_{0i})^2 - \ldots - (\sum_{i=0}^{n} \lambda_i\, e_{Ni})^2\right]\left[\| \bar{X} \|^2 - \sum_{j=0}^{N} (\bar{X},\, \bar{e}_j)^2 - \delta\right]$$

or

$$\sum_{i=0}^{n} [\lambda_i^2 \{t_i - (\bar{X}, \bar{e}_0)\, e_{0i} - \dots - (\bar{X}, \bar{e}_N)\, e_{Ni}\}^2$$

$$+ \; 2\lambda_i \sum_{\substack{k=0\\k>i}}^{n} \lambda_k \{t_i - (\bar{X}, \bar{e}_0)\, e_{0i} - \dots - (\bar{X}, \bar{e}_N)\, e_{Ni}\}\{t_k - (\bar{X}, \bar{e}_0)\, e_{0k} -$$

$$\dots - (\bar{X}, \bar{e}_N)\, e_{Nk}\} \;] \geq \sum_{i=0}^{n} [\lambda_i^2 - \lambda_i^2\, e_{0i}^2 - 2\lambda_i \sum_{\substack{k=0\\k>i}}^{n} \lambda_k\, e_{0i}\, e_{0k} - \dots$$

$$- \lambda_i^2\, e_{Ni}^2 \; - \; 2\lambda_i \sum_{\substack{k=0\\k>i}}^{n} \lambda_k\, e_{Ni}\, e_{Nk} \;] \cdot [\, \| \bar{X} \|^2 - \sum_{j=0}^{N} (\bar{X}, \bar{e}_j)^2 - \delta \,],$$

or

$$\sum_{i=0}^{n} [\lambda_i^2 \{ \{t_i - (\bar{X}, \bar{e}_0)\, e_{0i} - \dots - (\bar{X}, \bar{e}_N)\, e_{Ni} \}^2 - \{ \| \bar{X} \|^2 -$$

$$\sum_{j=0}^{N} (\bar{X}, \bar{e}_j)^2 - \delta\} \; + \; e_{0i}^2 \{ \| \bar{X} \|^2 - \sum_{j=0}^{N} (\bar{X}, \bar{e}_j)^2 - \delta\} + \dots +$$

$$e_{Ni}^2 \{ \| \bar{X} \|^2 - \sum_{j=0}^{N} (\bar{X}, \bar{e}_j)^2 - \delta\} \} \; + \; \lambda_i \{ 2 \sum_{\substack{k=0\\k>i}}^{n} \lambda_k \{t_i - (\bar{X}, \bar{e}_0)\, e_{0i}$$

$$- \dots - (\bar{X}, \bar{e}_N)\, e_{Ni}\} \{t_k - (\bar{X}, \bar{e}_0)\, e_{0k} - \dots - (\bar{X}, \bar{e}_N)\, e_{Nk}\} \; +$$

$$2 \{ \| \bar{X} \|^2 - \sum_{j=0}^{N} (\bar{X}, \bar{e}_j)^2 - \delta\} \{ \sum_{\substack{k=0\\k>i}}^{n} \lambda_k\, e_{0i}\, e_{0k} + \dots + \sum_{\substack{k=0\\k>i}}^{n} \lambda_k\, e_{Ni}\, e_{Nk} \} \} \,]$$

$$\geq 0.$$

If we let

$$A_i = \Big\{ (t_i - (\overline{X}, \overline{e}_0)\, e_{0i} - \dots - (\overline{X}, \overline{e}_N)\, e_{Ni})^2 - \{ \| \overline{X} \|^2 - \sum_{j=0}^{N} (\overline{X}, \overline{e}_j)^2 - \delta \}$$

$$+\ e_{0i}^2 \{ \| \overline{X} \|^2 - \sum_{j=0}^{N} (\overline{X}, \overline{e}_j)^2 - \delta \} + - - - + e_{Ni}^2 \{ \| \overline{X} \|^2 - \sum_{j=0}^{N} (\overline{X}, \overline{e}_j)^2 -$$

$\delta \} \Big\}$, and

$$B_i = 2 \Big\{ \sum_{\substack{k=0 \\ k>i}}^{n} \lambda_k \{ t_i - (\overline{X}, \overline{e}_0)\, e_{0i} - \dots - (\overline{X}, \overline{e}_N)\, e_{Ni} \} \{ t_k - (\overline{X}, \overline{e}_0)\, e_{0k} -$$

$$\dots - (\overline{X}, \overline{e}_N)\, e_{Nk}) + \{ \| \overline{X} \|^2 - \sum_{j=0}^{N} (\overline{X}, \overline{e}_j)^2 - \delta \} \{ \sum_{\substack{k=0 \\ k>i}}^{n} \lambda_k\, e_{0i}\, e_{0k} +$$

$$\dots + \sum_{\substack{k=0 \\ k>i}}^{n} \lambda_k\, e_{Ni}\, e_{Nk} \} \Big\} ,$$

we can write this inequality as $\sum_{i=0}^{n} (A_i \lambda_i^2 + B_i \lambda_i) \geq 0$, and this inequality is satisfied if

$$A_i \lambda_i^2 + B_i \lambda_i = 0,$$

for $i = 0,1,\dots,n$. Notice that these conditions are much stronger than are necessary and we will need to examine some cases that might arise.

Case 1: If $B_i \neq 0$, for all i, $0 \leq i < n$,

choose $\lambda_i = \dfrac{-B_i}{A_i}$.

<u>Case 2</u>: If $A_i \geq 0$ and $B_i = 0$, for some i,

$0 \leq i \leq n$, choose $\lambda_i = 1$.

<u>Case 3</u>: If $A_i < 0$ and $B_i = 0$, for some i, $0 \leq i \leq n$, choose λ_i satisfying

$$\frac{(B_{i-1})^2}{\lambda_i^2} \geq 4A_{i-1} A_i.$$

Then $(B_{i-1})^2 - 4A_{i-1} A_i \lambda_i^2 \geq 0$, and we are assured of a solution λ_{i-1} to the equation

$$A_{i-1} \lambda_{i-1}^2 + B_{i-1} \lambda_{i-1} + A_i \lambda_i^2 = 0.$$

Notice that the left side of this equation is just the sum of the ith and (i-1)st terms of the sum $\sum_{i=0}^{n} (A_i \lambda_i^2 + B_i \lambda_i)$. Since $A_i < 0$, no difficulty is encountered in choosing λ_i to satisfy

$$\frac{(B_{i-1})^2}{\lambda_i^2} \geq 4A_{i-1} A_i,$$

if $A_{i-1} \geq 0$. If $A_{i-1} < 0$, then we must note that B_{i-1} is a function of λ_i, and we must be careful in the choice of λ_i.

But $\frac{(B_{i-1})^2}{\lambda_i^2}$ has a minimum value at its only critical point, hence we can choose λ_i such that $\frac{(B_{i-1})^2}{\lambda_i^2} \geq 4A_{i-1} A_i$.

III. REFERENCES

1. "Existence, Uniqueness and Derivation of Multivariable Approximating Functions", Annual Report, Contract Number NAS8-2642, Prepared for George C. Marshall Space Flight Center, Huntsville, Alabama.

2. Achieser, N. I., Theory of Approximation, Ungar Publishing Company.

ENGINEERING EXPERIMENT STATION
RICH ELECTRONIC COMPUTER CENTER
GEORGIA INSTITUTE OF TECHNOLOGY

LEAST SQUARES ESTIMATION OF REGRESSION COEFFICIENTS IN A SPECIAL CLASS OF POLYNOMIAL MODELS

By

D. G. Herr, J. J. Goode,
I. E. Perlin, and J. H. MacKay

ATLANTA, GEORGIA

ENGINEERING EXPERIMENT STATION
GEORGIA INSTITUTE OF TECHNOLOGY
ATLANTA, GEORGIA

LEAST SQUARES ESTIMATION OF REGRESSION COEFFICIENTS IN A SPECIAL CLASS OF POLYNOMIAL MODELS

By

D. G. Herr, J. J. Goode,
I. E. Perlin, and J. H. MacKay

SUMMARY

20962

A

A method is proposed for the least squares fitting of a polynomial to data with applications to fitting solutions of the guidance equations. This method depends on restricting the class of polynomials (balanced polynomials) as well as on solving the guidance equations at certain sets of points.

Author

I. INTRODUCTION

The problem with which we are concerned is that of approximating a real valued function of several real variables given a collection of points in the domain of the function and the corresponding values of the function at these points. Furthermore, we are considering a polynomial approximation of the function and are assuming the least squares criterion for the best approximation. Theoretically, then, our problem is easy -- simply use the polynomial of the chosen degree with the least squares estimates of the coefficients. However, from the practical point of view the problem is not so easy. Actually finding the least squares coefficients may be an almost impossible task when one is fitting a polynomial of several variables and modest degree. The inversion of the coefficient matrix of the normal equations is the usual problem.

The general methods for finding the least squares coefficients can be divided into two major categories--those which apply for arbitrarily chosen data points and those which depend on some special arrangement or design of the data points. The methods thus far proposed for arbitrarily chosen data points do not seem substantially to reduce the calculational difficulties from those of inverting the coefficient matrix of the normal equations. However, if one is willing to allow any apriori design of the data points, it is possible to have a design which will yield an easily invertable coefficient matrix. There is, of course, a middle ground between that of no restriction on the arrangement of data points (design) and that of the very severe restrictions needed to produce an easily invertable coefficient matrix. It is in this area of moderate restrictions on the design of the data points that we have had some success. We shall call our design of the data points a rectangular design. In the statistical literature this design is called a factorial design.

By using a rectangular design and a special form of polynomial called a balanced polynomial we have been able to calculate the least squares coefficients with a considerable reduction in calculational difficulty in the sense that several lower order matrices are easier to invert than one of higher order. The process by which we calculate the least squares coefficients will be called the step procedure.

II. RECTANGULAR DESIGNS

Suppose the domain of the function to be approximated is a subset of π-dimensional Euclidian space. Let $(x^{(1)}, \dots, x^{(\pi)})$ be a typical point and define

$$D_i = \left\{ x_{t_i}^{(i)} : t_i = 1, \dots, T_i , \; x_{t_i} \neq x_{s_i} \text{ if } t_i \neq s_i \right\} .$$

Then the cartesian product

$$D_1 \text{ x } D_2 \text{ x } \dots \text{ x } D_\pi = D$$

will be a subset of π-dimensional Euclidian space. We define a rectangular design to be any such D . Note that the T_i's need not be equal and the $x_{t_i}^{(i)}$ need not be equally spaced.

Step Procedure: The step procedure is most easily explained by an example. Let us consider a function of two variables, f , and consider an approximation of f by means of a second degree polynomial. Denoting $f(u,v)$ by y we have

$$y \approx (a_{11} + a_{21}u + a_{31}u^2) + (a_{12} + a_{22}u)v + a_{13}v^2 .$$

Suppose the data is in a rectangular design, say

$$D = D_1 \text{ x } D_2 , \; D_1 = (u_1 , \dots , u_n) , \; D_2 = (v_1 , \dots , v_m)$$

then we may use the step procedure to find estimates, not necessarily the best, of the a's . The procedure is as follows:

1. Hold u fixed at say u_i and define b_{i1} , b_{i2} , b_{i3} by

$$b_{i1} = a_{11} + a_{21}u_i + a_{31}u_i^2$$

$$b_{i2} = a_{12} + a_{22}u_i$$

$$b_{i3} = a_{13}$$

and consider

$$y_{ij} \approx b_{i1} + b_{i2}v_j + b_{i3}v_j^2$$

2. For each fixed i find the least squares estimates of b_{i1} , b_{i2} , b_{i3} .

3. Using these estimates as if they were observed values of $a_{11} + a_{21}u_i + a_{22}u_i^2$, $a_{12} + a_{22}u_i$, a_{13} respectively find the least squares estimates of a_{11} , a_{21} , a_{22} ; a_{12} , a_{22} ; and a_{13} .

Note that instead of a 6×6 matrix inversion as in the case of finding direct least squares estimates of the a's , we were only required to invert several smaller matrices of maximum size 3×3 . We could also have written the polynomial approximation as

$$y \approx (a_{11} + a_{12}v + a_{13}v^2) + (a_{21} + a_{22}v)u + a_{31}u^2$$

and used the procedure just as well. The estimates of the a's in this case would, in general, be different from those found above.

It is not difficult to show that in a general n^{th} degree polynomial of π variables the estimate of the coefficient of the highest power of the variable which appears in the first step of the step procedure is indeed the least squares estimate. We shall denote this result as theorem 1 . In general the estimates of the other coefficients do not have this property.

III. BALANCED POLYNOMIALS

As motivation for considering balanced polynomials, think of expanding a function of π variables, $x^{(1)}, \ldots, x^{(\pi)}$, in a power series in $x^{(\pi)}$ and approximate this by the first $L_\pi + 1$ terms; i.e., a polynomial in $x^{(\pi)}$ of degree L_π. Now expand the coefficients of this polynomial in power series in $x^{(\pi-1)}$ and approximate these series by their first $L_{\pi-1} + 1$ terms. Continue this process until all the variables have been used. Note this yields a polynomial in $x^{(1)}, \ldots, x^{(\pi)}$ of degree $L_1 L_2 \ldots L_\pi$ which is not the general polynomial of this degree. For example, if $\pi = 2$, $L_1 = L_2 = 2$ we have the balanced polynomial

$$(a_{11} + a_{21}u + a_{31}u^2) + (a_{12} + a_{22}u + a_{32}u^2)v + (a_{13} + a_{23}u + a_{33}u^2)v^2 .$$

This polynomial is a fourth degree polynomial in u, v but the u^4, u^3, v^4, v^3, u^3v, v^3u terms are missing. Notice, however, that all the terms of the general second degree polynomial are present. So if higher degree terms are not objectionable, it would seem that if a general polynomial in π variables of degree L provides a reasonable approximation, a balanced polynomial in π variables with $\min_{j=1,\ldots,\pi} L_j \geq L$ would give an even better approximation.

In general a polynomial of the form

$$\sum_{\ell_1=1}^{L_1+1} \ldots \sum_{\ell_\pi=1}^{L_\pi+1} a_{\ell_1 \ldots \ell_\pi} \; x^{(1)}_{\ell_1} \ldots x^{(\pi)}_{\ell_\pi} , \quad x^{(i)}_{\ell_i} = (x^{(i)})^{\ell_i - 1}$$

will be called a balanced polynomial. We show in theorem 2 that the step procedure applied to a balanced polynomial over a rectangular design will yield the least squares estimates of all the coefficients.

IV. PROPERTIES OF RECTANGULAR DESIGN AND BALANCED POLYNOMIALS

Consider the general d^{th} degree polynomial in the π variables $x^{(1)}, \ldots, x^{(\pi)}$ which we shall write as

$$(1) \quad a_{11\ldots1} + \ldots + a_{1\ldots i\ldots1}x_1^{(i)} + \ldots + (\text{terms in } x^{(1)}, \ldots, x^{(\pi)} \text{ of degree } \leq d) + a_{1\ldots1d+1}(x^{(\pi)})^d \ .$$

We shall call $x^{(\pi)}$ the leading variable. Clearly this general polynomial may be written with any $x^{(i)}$ as the leading variable but in what follows we shall be concerned with the specific form of the polynomial in (1) and thus the leading variable will be $x^{(\pi)}$. If we use such a polynomial to approximate a real valued function f of π variables $x^{(1)}, \ldots, x^{(\pi)}$; we have the following result.

<u>THEOREM 1</u>: In the case of a general d^{th} degree polynomial in π variables the step procedure over a rectangular design yields the same estimate for the coefficient of the d^{th} power of the leading variable as the least squares procedure over the same design.

Before presenting a proof of theorem 1 we shall exhibit an example which shows that theorem 1 is best possible in the sense that in general the step procedure estimates and the least squares estimates of the other coefficients do not agree. In particular this will justify the use of the specific form of (1) and the "leading variable" terminology.

Consider the general second degree polynomial in two variables

$$a_{11} + a_{21}u + a_{31}u^2 + (a_{12} + a_{22}u)v + a_{13}v^2$$

as an approximation of a real valued function f of two real variables u,v . Let $y_{ij} = f(u_i , v_j)$ and thus suppose the expected value of y_{ij} given by

$$E(y_{ij}) = a_{11} + a_{21}u_i + a_{31}u_i^2 + (a_{12} + a_{22}u_i)v_j + a_{13}v_j^2$$

or in vector-matrix notation

$$E(\underline{y}) = X\,\underline{a}$$

where

$$\underline{y} = \begin{pmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \\ y_{23} \\ y_{31} \\ y_{32} \\ y_{33} \end{pmatrix} \qquad X = \begin{pmatrix} 1 & u_1 & u_1^2 & u_1v_1 & v_1 & v_1^2 \\ 1 & u_1 & u_1^2 & u_1v_2 & v_2 & v_2^2 \\ 1 & u_1 & u_1^2 & u_1v_3 & v_3 & v_3^2 \\ 1 & u_2 & u_2^2 & u_2v_1 & v_1 & v_1^2 \\ 1 & u_2 & u_2^2 & u_2v_2 & v_2 & v_2^2 \\ 1 & u_2 & u_2^2 & u_2v_3 & v_3 & v_3^2 \\ 1 & u_3 & u_3^2 & u_3v_1 & v_1 & v_1^2 \\ 1 & u_3 & u_3^2 & u_3v_2 & v_2 & v_2^2 \\ 1 & u_3 & u_3^2 & u_3v_3 & v_3 & v_3^2 \end{pmatrix}$$

$$\underline{a} = \begin{pmatrix} a_{11} \\ a_{21} \\ a_{31} \\ a_{12} \\ a_{22} \\ a_{13} \end{pmatrix} \qquad \text{for } i,j = 1, 2, 3 \ .$$

The least squares estimates of the coefficients may be found by solving the normal equations [1]

$$X'\,X\,\underline{a} = X'\,\underline{y} \quad .$$

If in our example we consider the rectangular design

$$D = D_1 \times D_2 \quad ; \quad D_1 = \left\{ -1,\ 0,\ 1 \right\} \qquad D_2 = \left\{ -2,\ 0,\ 1 \right\}$$

the normal equations become

$$\begin{pmatrix} 9 & 0 & 6 & 0 & -3 & 15 \\ 0 & 6 & 0 & -2 & 0 & 0 \\ 6 & 0 & 6 & 0 & -2 & 10 \\ 0 & -2 & 0 & 10 & 0 & 0 \\ -3 & 0 & -2 & 0 & 15 & -21 \\ 15 & 0 & 10 & 0 & -21 & 51 \end{pmatrix} \begin{pmatrix} a_{11} \\ a_{21} \\ a_{31} \\ a_{12} \\ a_{22} \\ a_{13} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ -1 & -1 & -1 & 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 \\ 2 & 0 & -1 & 0 & 0 & 0 & -2 & 0 & 1 \\ -2 & 0 & 1 & -2 & 0 & 1 & -2 & 0 & 1 \\ 4 & 0 & 1 & 4 & 0 & 1 & 4 & 0 & 1 \end{pmatrix} \begin{pmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \\ y_{23} \\ y_{31} \\ y_{32} \\ y_{33} \end{pmatrix}$$

The solution of this system is

$$a_{11} = \frac{1}{54} \left(48 \sum_i y_{i1} + 30 \sum_i y_{i2} + 21 \sum_i y_{i3} - 18 \sum_j y_{1j} - 18 \sum_j y_{3j} \right)$$

$$a_{21} = \frac{1}{28} \left(-5 \sum_j y_{1j} + 5 \sum_j y_{3j} + 2 y_{11} - y_{13} - 2y_{31} + y_{33} \right)$$

$$a_{31} = \frac{1}{54} \left(-18 \sum_i y_{i1} - 18 \sum_i y_{i2} - 18 \sum_i y_{i3} + 27 \sum_j y_{1j} + 27 \sum_j y_{3j} \right)$$

$$a_{12} = \frac{1}{28} \left(- \sum_j y_{1j} + \sum_j y_{3j} + 6\, y_{11} - 3\, y_{13} - 6\, y_{31} + 3\, y_{33} \right)$$

$$a_{22} = \frac{1}{54} \left(-3 \sum_i y_{i1} - 9 \sum_i y_{i2} + 12 \sum_i y_{i3} \right)$$

$$a_{13} = \frac{1}{54} \left(3 \sum_i y_{i1} - 9 \sum_i y_{i2} + 6 \sum_i y_{i3} \right)$$

Now consider the same design and use the step procedure to estimate the coefficients. Thus, write the polynomial as

$$b_1 + b_2 v + b_3 v^2$$

where

$$b_1 = a_{11} + a_{21}u + a_{31}u^2$$

$$b_2 = a_{12} + a_{22}u$$

$$b_3 = a_{13} \quad .$$

For fixed i find the least squares estimates of $b_{1i} = a_{11} + a_{21}u_i + a_{31}u_i^2$, $b_{2i} = a_{12} + a_{22}u_i$, $b_{3i} = a_{13}$. We obtain the normal equations in v alone:

$$(V' V) \underline{b}_i = V' \underline{y}_i$$

where

$$V = \begin{pmatrix} 1 & v_1 & v_1^2 \\ 1 & v_2 & v_2^2 \\ 1 & v_3 & v_3^2 \end{pmatrix} , \quad \underline{b}_i = \begin{pmatrix} b_{1i} \\ b_{2i} \\ b_{3i} \end{pmatrix} , \quad \underline{y}_i = \begin{pmatrix} y_{i1} \\ y_{i2} \\ y_{i3} \end{pmatrix} \quad .$$

The solution of this system is

$$b_{1i} = y_{i2}$$

$$b_{2i} = -\frac{1}{6} y_{i1} - \frac{1}{2} y_{i2} + \frac{2}{3} y_{i3}$$

$$b_{3i} = \frac{1}{6} y_{i1} - \frac{1}{2} y_{i2} + \frac{1}{3} y_{i3} \quad . \quad (i = 1, 2, 3)$$

The second step is to treat the u's as observations on the polynomials $a_{11} + a_{21}u + a_{31}u^2$, $a_{21} + a_{22}u$, a_{13} and find the least squares estimates of the a's . For b_1 the normal equations are

$$(U_1' U_1) \underline{a}_1 = U_1' \underline{b}_1$$

where

$$U_1 = \begin{pmatrix} 1 & u_1 & u_1^2 \\ 1 & u_2 & u_2^2 \\ 1 & u_3 & u_3^2 \end{pmatrix} , \quad \underline{a}_1 = \begin{pmatrix} a_{11} \\ a_{21} \\ a_{31} \end{pmatrix} , \quad \underline{b}_1 = \begin{pmatrix} b_{11} \\ b_{12} \\ b_{13} \end{pmatrix} \quad .$$

The solution of this system is

$$\begin{aligned} a_{11} &= b_{12} && = y_{22} \\ a_{21} &= -\tfrac{1}{2}(b_{11} + b_{13}) && = -\tfrac{1}{2}(y_{12} + y_{32}) \\ a_{31} &= \tfrac{1}{2}(b_{11} + b_{13}) - b_{12} && = \tfrac{1}{2}(y_{12} + y_{32}) - y_{22} \ . \end{aligned}$$

In the case of b_2 the normal equations are

$$(U_2' U_2)\,\underline{a}_2 = U_2'\,\underline{b}_2 \ ,$$

where

$$U_2 = \begin{pmatrix} 1 & u_1 \\ 1 & u_2 \\ 1 & u_3 \end{pmatrix} \ , \qquad \underline{a}_2 = \begin{pmatrix} a_{12} \\ a_{22} \end{pmatrix} \ , \qquad \underline{b}_2 = \begin{pmatrix} b_{21} \\ b_{22} \\ b_{23} \end{pmatrix} \ .$$

The solution of this system is

$$a_{12} = \frac{1}{3} \sum_{i=1}^{3} b_{1i} = \frac{1}{54}\left(-3\,\Sigma\, y_{i1} - 27\,\Sigma\, y_{i2} + 36\,\Sigma\, y_{i3} \right)$$

$$a_{22} = \frac{1}{2}\left(b_{13} - b_{11} \right) = \frac{1}{12}\left(-y_{31} - 3y_{32} + 4y_{33} + y_{11} + 3y_{12} - 4y_{13} \right) \ .$$

Note at this point that none of the step procedure estimates agrees with the least squares estimate.

Finally consider b_3 and the normal equations

$$(U_3' U_3)\,\underline{a}_3 = U_3'\,\underline{b}_3$$

where

$$U_3 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \ , \qquad \underline{a}_3 = a_{13} \ , \qquad \underline{b}_3 = \begin{pmatrix} b_{31} \\ b_{32} \\ b_{33} \end{pmatrix}$$

so that

$$a_{13} = \frac{\sum_{j=1}^{3} b_{3j}}{3} = \frac{1}{54}\left(3\,\Sigma\, y_{i1} - 9\,\Sigma\, y_{i2} + 6\,\Sigma\, y_{i3} \right)$$

which does agree with the least squares estimate of a_{13} .

Thus, we see that the step procedure for estimating the coefficients of a general polynomial over a rectangular design is not equivalent to least squares estimation over the same design. However, in theorem 2 we shall give conditions sufficient for the equivalence of the two procedures. We now present a proof of theorem 1 in the case $d = 2$, $\pi = 2$. (For the general proof see Appendix B.)

Consider the rectangular design

$$D = D_1 \times D_2 \quad ; \quad D_1 = \left\{ u_1 , u_2 , u_3 \right\} \quad ; \quad D_2 = \left\{ v_1 , v_2 , v_3 \right\}$$

and the polynomial

$$(*) \qquad (a_{11} + a_{21}u + a_{31}u^2) + (a_{12} + a_{22}u)v + a_{13}v^2 \quad .$$

written in preparation for the first step of the step procedure as

$$b_1 + b_2 v + b_3 v^2 \quad .$$

where $b_1 = a_{11} + a_{21}u + a_{31}u^2$, $b_2 = a_{12} + a_{22}u$, $b_3 = a_{13}$. Let $y_{ij} = f(u_i v_j)$ where f is the function to be approximated by the polynomial (*)

If we can demonstrate that the step procedure estimate of $a_{13} = b_3$ is a linear combination of the components of $X'\underline{y}$, using the notation of the example, and show that such an estimate is unbiased; then the step procedure estimate is the least squares estimate. (See Appendix A)

Since the first step of the step procedure is a least squares estimation, the step procedure estimate of b_3 , $\hat{b}_3$, is unbiased. Furthermore,

$$\begin{pmatrix} \hat{b}_{1i} \\ \hat{b}_{2i} \\ \hat{b}_{3i} \end{pmatrix} = (V'V)^{-1} \begin{pmatrix} \sum_j y_{ij} \\ \sum_j v_j y_{ij} \\ \sum_j v_j^2 y_{ij} \end{pmatrix}$$

that is, $\hat{b}_{3i}$ is a linear combination of

$$\sum_j y_{ij} \quad , \quad \sum_j v_j y_{ij} \quad , \quad \sum_j v_j^2 y_{ij}$$

for each i . Since $b_{3i} = a_{13}$ for each i the second step of the step procedure gives

$$\frac{\sum_{i=1}^{3} \hat{b}_{3i}}{3}$$

as the step procedure estimate of a_{13} . Clearly this is unbiased if $\hat{b}_{3i}$ is and this estimate is a linear combination of

$$\sum_i \sum_j y_{ij} \quad , \quad \sum_i \sum_j v_j y_{ij} \quad , \quad \sum_i \sum_j v_j^2 y_{ij} \quad .$$

However, the components of $X'\underline{y}$ are

$$\sum_i \sum_j y_{ij} \quad , \quad \sum_{ij} u_i y_{ij} \quad , \quad \sum_{ij} u_i^2 y_{ij} \quad , \quad \sum_{ij} u_i v_j y_{ij} \quad , \quad \sum_{ij} v_j y_{ij} \quad , \quad \sum_{ij} v_j^2 y_{ij}$$

so that the step procedure estimate of a_{13} is a linear combination of these components, specifically of the first, fifth and sixth. Thus, the proof is complete for this special case.

If we are willing to restrict ourselves to balanced polynomials, we may use the following result.

THEOREM 2: The step procedure when applied to a balanced polynomial approximation of a real function of several real variables over a rectangular design will yield the least squares estimates of the coefficients.

Consider the special case of a balanced polynomial in two variables each with maximum degree 2 ,

$$(a_{11} + a_{21}u + a_{31}u^2) + (a_{12} + a_{22}u + a_{32}u^2)\, v + (a_{13} + a_{23}u + a_{33}u^2)\, v^2 ,$$

as an approximation of a real function f of two real variables u,v over the rectangular design

$$D = D_1 \times D_2 , \quad D_1 = \left\{ u_1 , u_2 , u_3 \right\} , \quad D_2 = \left\{ v_1 , v_2 , v_3 \right\} .$$

Let $y_{t_1t_2} = f\,(u_{t_1} , v_{t_2})$; $t_1 = 1, 2, 3$; $t_2 = 1, 2, 3$.

First we shall consider the least squares criterion for estimates of the a's and generate the normal equations ; then we shall show that the step procedure estimates of the a's satisfy the normal equations and are, therefore, least squares estimates.

Define S by

$$S = \sum_{t_1=1}^{3} \sum_{t_2=1}^{3} \left\{ y_{t_1t_2} - \sum_{\ell_2=1}^{3} \sum_{\ell_1=1}^{3} a_{\ell_1\ell_2} u_{t_1}^{\ell_1-1} v_{t_2}^{\ell_2-1} \right\}^2$$

and calculate $\frac{\partial}{\partial a_{\alpha_1\alpha_2}} S$. Setting this partial derivative equal to zero, we arrive at the equation

$$\sum_{t_1=1}^{3} \sum_{t_2=1}^{3} y_{t_1 t_2} u_{t_1}^{\alpha_1 - 1} v_{t_2}^{\alpha_2 - 1} = \sum_{\ell_2=1}^{3} \sum_{\ell_1=1}^{3} a_{\ell_1 \ell_2} \left\{ \sum_{t_1=1}^{3} \sum_{t_2=1}^{3} u_{t_1}^{\ell_1 - 1} u_{t_1}^{\alpha_1 - 1} v_{t_2}^{\ell_2 - 1} v_{t_2}^{\alpha_2 - 1} \right\} .$$

Now employing the properties of the rectangular design we have

$$(4) \quad \sum_{t_1=1}^{3} \sum_{t_2=1}^{3} y_{t_1 t_2} u_{t_1}^{\alpha_1 - 1} v_{t_2}^{\alpha_2 - 1} = \sum_{\ell_1=1}^{3} \sum_{\ell_2=1}^{3} a_{\ell_1 \ell_2} \left\{ \left(\sum_{t_1=1}^{3} u_{t_1}^{\ell_1 - 1} u_{t_1}^{\alpha_1 - 1} \right) \left(\sum_{t_2=1}^{3} v_{t_2}^{\ell_2 - 1} v_{t_2}^{\alpha_1 - 1} \right) \right\} .$$

We shall define the matrices U,V as follows

$$U = \begin{pmatrix} 1 & u_1 & u_1^2 \\ 1 & u_2 & u_2^2 \\ 1 & u_3 & u_3^2 \end{pmatrix} \qquad V = \begin{pmatrix} 1 & v_1 & v_1^2 \\ 1 & v_2 & v_2^2 \\ 1 & v_3 & v_3^2 \end{pmatrix}$$

Then clearly $\sum_{t_1=1}^{3} u_{t_1}^{\ell_1 - 1} u_{t_1}^{\alpha_1 - 1}$ is in the ℓ_1, α_1 position of the matrix $U'U$.

Similarly for $\sum_{t_2=1}^{3} v_{t_2}^{\ell_2 - 1} v_{t_2}^{\alpha_2 - 1}$. Thus define

$$U_{\ell_1 \alpha_1} = \sum_{t_1=1}^{3} u_{t_1}^{\ell_1 - 1} u_{t_1}^{\alpha_1 - 1}$$

$$V_{\ell_2 \alpha_2} = \sum_{t_2=1}^{3} v_{t_2}^{\ell_2 - 1} v_{t_2}^{\alpha_2 - 1}$$

and (4) becomes

$$(5) \quad \sum_{t_1=1}^{3} \sum_{t_2=1}^{3} y_{t_1 t_2} u_{t_1}^{\alpha_1 - 1} v_{t_2}^{\alpha_2 - 1} = \sum_{\ell_1=1}^{3} \sum_{\ell_2=1}^{3} A_{\ell_1 \ell_2} U_{\ell_1 \alpha_1} V_{\ell_2 \alpha_2}$$

Equation (5) is a typical equation from the set of normal equations.

We shall now use the step procedure to estimate a coefficient, $A_{s_1 s_2}$. In order to facilitate the writing down of this estimate, we shall have need of the following notation. Let

$$(U^{\ell_1 s_1}) = (U'U)^{-1} \qquad \ell_1 = 1, 2, 3 \quad s_1 = 1, 2, 3$$

$$(V^{\ell_2 s_2}) = (V'V)^{-1} \qquad \ell_2 = 1, 2, 3 \quad s_2 = 1, 2, 3$$

and note $u_{r_1}^{\ell_1 -1}$ is in the ℓ_1, r_1 position of U' and similarly for $v_{r_2}^{\ell_2 -1}$.

The first step of the step procedure for finding an estimate of $A_{s_1 s_2}$ is

$$A^{(1)}_{t_1 s_2} = (V'V)^{-1} V' \underline{Y}_{t_1}$$

$$= \sum_{r_2=1}^{3} \left\{ \sum_{\ell_2=1}^{3} V^{\ell_2 s_2} v_{r_2}^{\ell_2 -1} \right\} y_{t_1 r_2}$$

where $\underline{Y}_{t_1} = (y_{t_1 1} , y_{t_1 2} , y_{t_1 3})$. The second and in this case final step is then

$$A_{s_1 s_2} = (U'U)^{-1} U' a^{(1)}_{s_2}$$

$$= \sum_{r_1=1}^{3} \left\{ \sum_{\ell_1=1}^{3} U^{\ell_1 s_1} u_{r_1}^{\ell_1 -1} \right\} a^{(1)}_{r_1 s_2}$$

$$= \sum_{r_1=1}^{3} \quad \sum_{\ell_1=1}^{3} U^{\ell_1 s_1} u_{r_1}^{\ell_1 -1} \sum_{r_1=1}^{3} \quad \sum_{\ell_2=1}^{3} V^{\ell_2 s_2} v_{r_2}^{\ell_2 -1} y_{r_1 r_2}$$

Using the fact that we have a balanced polynomial over a rectangular design we may write

$$A_{s_1 s_2} = \sum_{r_1=1}^{3} \sum_{r_2=1}^{3} \left\{ y_{r_1 r_2} \sum_{\ell_1=1}^{3} \sum_{\ell_2=1}^{3} U^{\ell_1 s_1} V^{\ell_2 s_2} u_{r_1}^{\ell_1 -1} v_{r_2}^{\ell_2 -1} \right\} .$$

If we substitute $A_{s_1 s_2}$ for $A_{\ell_1 \ell_2}$ in equation (5), the right hand side becomes

$$\sum_{s_1=1}^{3} \sum_{s_2=1}^{3} A_{s_1 s_2} U_{s_1 \alpha_1} V_{s_2 \alpha_2} =$$

$$\sum_{r_1=1}^{3} \sum_{r_2=1}^{3} y_{r_1 r_2} \left\{ \sum_{\ell_1=1}^{3} \sum_{\ell_2=1}^{3} [u_{r_1}^{\ell_1 - 1} v_{r_2}^{\ell_2 - 1} (\sum_{s_1=1}^{3} U^{\ell_1 s_1} U_{s_1 \alpha_1})(\sum_{s_2=1}^{3} V^{\ell_2 s_2} V_{s_2 \alpha_2})] \right\} .$$

However $\sum_{s_1=1}^{3} U^{\ell_1 s_1} U_{\ell_1 \alpha_1} = \delta_{\ell_1 \alpha_1} = 0$ or 1 depending on whether $\ell_1 \neq \alpha_1$ or $\ell_1 = \alpha_1$. Similarly, $\sum_{s_2=1}^{3} V^{\ell_2 s_2} V_{s_2 \alpha_2} = \delta_{\ell_2 \alpha_2}$. So that we have the right hand side of (5) equal to

$$\sum_{r_1=1}^{3} \sum_{r_2=1}^{3} y_{r_1 r_2} \left\{ \sum_{\ell_1=1}^{3} \sum_{\ell_2=1}^{3} u_{r_1}^{\ell_1 - 1} v_{r_2}^{\ell_2 - 1} \delta_{\ell_1 \alpha_1} \delta_{\ell_2 \alpha_2} \right\} =$$

$$\sum_{r_1=1}^{3} \sum_{r_2=1}^{3} y_{r_1 r_2} u_{r_1}^{\alpha_1 - 1} v_{r_2}^{\alpha_2 - 1}$$

which is the left hand side of equation (5) . Thus $A_{s_1 s_2}$ is a solution of the normal equations and the proof of theorem 2 is complete for this special case.

V. IMPLICATIONS AND EXTENSIONS

Comparison to ANOVA

The analysis of variance model for a factorial design which includes all of the interaction terms is equivalent to a balanced polynomial model in which the degree of the polynomial in a given variable is one less than the number of levels of the factor corresponding to that variable. In the analysis of variance model we break up the degrees of freedom for a factor into each of the different levels and in a polynomial model we use the constant, linear, and quadratic parts. If we have a factor at levels a, b, and c then we may think of these three degrees of freedom as corresponding to the space spanned by

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} .$$

Equivalently, we may consider the space spanned by

$$\begin{pmatrix} 1 & a & a^2 \\ 1 & b & b^2 \\ 1 & c & c^2 \end{pmatrix} .$$

The first is the analysis of variance model and the second is the polynomial model.

A factorial design in which all interactions above order d are assumed to be zero is equivalent to a polynomial model in which cross products involving more than $d + 1$ factor are omitted.

Relaxation of Balanced Polynomial Conditions

We have seen in theorem 1 that the rectangular design enables us to use the very easy step procedure to find the least squares estimate of the coefficients of the highest power of each variable in a model which is the general polynomial of degree d . In theorem 2 we see that the rectangular design enables us to use the step procedure to calculate the least squares estimates of all of the coefficients of a model which is a balanced polynomial. We may now ask; is it necessary to have a balanced polynomial to get all of the coefficients by the step procedure? Is it possible to have other polynomial models in which the step procedure gives the least squares estimates of some terms other than just the highest power?

To gain some insight into these questions we shall consider as an example the two factor model

$$E\, y_{ij} = P\,(u_i\,,\, v_j)$$

where P is a polynomial in u and v and the design is a rectangular design in which u has 4 values and v has 3 .

Now we apply the step procedure with leading variable v . We write $P\,(u,\, v)$ as a polynomial in v .

$$E\, y_{ij} = P_o(u_i) + v_j P_1(u_i) + v_j^2 P_2(u_i) \quad .$$

Let $v = \begin{pmatrix} 1 & v_1 & v_1^2 \\ 1 & v_2 & v_2^2 \\ 1 & v_3 & v_3^2 \end{pmatrix}$. Let $y_i = \begin{pmatrix} y_{i1} \\ y_{i2} \\ y_{i3} \end{pmatrix}$.

Then the estimates of P_o , P_1 , and P_2 are given by

$$\begin{pmatrix} \hat{P}_o\ (u_i) \\ \hat{P}_1\ (u_i) \\ \hat{P}_2\ (u_i) \end{pmatrix} = (v^T v)^{-1} v^T\ y_i \quad .$$

In particular $\hat{P}_\ell\ (u_i) = L_\ell \begin{pmatrix} \sum_j y_{ij} \\ \sum_j v_j y_{ij} \\ \sum_j v_j^2 y_{ij} \end{pmatrix}$ where L_ℓ stands for some linear combination. If we assume that $P_o(u_i) = a_{oo} + a_{1o}u + a_{2o}u^2$, $P_1(u) = a_{o1} + a_{11}u$, and $P_2(u) = a_{o2}$ then we estimate a_{o2} by averaging $\hat{P}_2(u_i)$ over the values of u . That is

$$\hat{a}_{o2} = L_{o2} \begin{pmatrix} \sum_{ij} y_{ij} \\ \sum_{ij} v_j\ y_{ij} \\ \sum_{ij} v_j^2\ y_{ij} \end{pmatrix} \quad .$$

Now this is a least squares estimate of a_{o2} only if it is a linear combination of the right side of the least squares normal equations. That is, only if it is a linear combination of

$$x^T y = \begin{pmatrix} \sum_{ij} y_{ij} \\ \sum_{ij} u_i\ y_{ij} \\ \sum_{ij} u_i^2\ y_{ij} \\ \sum_{ij} v_j\ y_{ij} \\ \sum_{ij} u_i\ v_j\ y_{ij} \\ \sum_{ij} v_j^2\ y_{ij} \end{pmatrix} \quad .$$

$\hat{a}_{o2}$ is a linear combination of these terms.

We estimate a_{01} and a_{11} by

$$\begin{pmatrix} \hat{a}_{01} \\ \hat{a}_{11} \end{pmatrix} = \begin{pmatrix} 4 & \Sigma u_i \\ \Sigma u_i & \Sigma u_i^2 \end{pmatrix}^{-1} \begin{pmatrix} 1 & 1 & 1 & 1 \\ u_1 & u_2 & u_3 & u_4 \end{pmatrix} \begin{pmatrix} P_i\ (u_1) \\ P_i\ (u_2) \\ P_i\ (u_3) \\ P_i\ (u_4) \end{pmatrix}$$

$$= L \begin{pmatrix} \sum\limits_{ij} y_{ij} \\ \sum\limits_{ij} v_j\ y_{ij} \\ \sum\limits_{ij} v_j^2\ y_{ij} \\ \sum\limits_{ij} u_i\ y_{ij} \\ \sum\limits_{ij} u_i\ v_j\ y_{ij} \\ \sum\limits_{ij} u_i\ v_j^2\ y_{ij} \end{pmatrix} .$$

All of the components of this vector except the last one are in $x^T y$. Hence $\hat{a}_{01}$ or $\hat{a}_{11}$ are not least squares estimates unless the data points u_i , v_j are such that the linear combinations symbolized by L do not involve this last term.

Now we could also put $a_{12}uv^2$ in the model so as to put $\sum\limits_{ij} u_i v_j^2 y_{ij}$ in the right side of the least squares normal equations. By continually putting terms in the model as needed in this example we find that to determine the least squares estimates of all of the coefficients by the step procedure independent of the data points (except that the design be rectangular) it is necessary that the polynomial in the model be balanced. This example also indicates how we would go about expanding the polynomial model so as to estimate certain coefficients by the step procedure. Having estimated some of the coefficients, we may eliminate them from y and do an ordinary least squares regression, if it is then practicable.

APPENDIX A: STATISTICAL BACKGROUND

It is assumed that the reader is familiar with such terms as expected value, random variable, variance, etc. If not, ready reference to these terms may be found in such books as Cramer [2] and Loeve [3] .

We shall be concerned here with independent random variables $y_1, \ldots, y_n$ such that the expected value of y_i is a linear function of m parameters $p_1, \ldots, p_m$ and the variance of y_i is σ^2 , i.e.,

$$E(\underline{y}) = A\,\underline{p}$$

$$\mathrm{Var}(y_i) = \sigma^2 \qquad i = 1, \ldots, n$$

where $\underline{y}' = (y_1, \ldots, y_n)$, $\underline{p}' = (p_1, \ldots, p_n)$ and $A = (a_{ij})$ is a known real $n \times m$ matrix. We shall be interested in estimating by functions of $y_1, \ldots, y_n$ certain linear functions of the parameters, say $\underline{\ell}'\underline{p}$ where $\underline{\ell}' = (\ell_1, \ldots, \ell_m)$. We call an estimate of $\underline{\ell}'\underline{p}$ which has expected value $\underline{\ell}'\underline{p}$ an unbiased estimate. If the estimate is also a linear function of the y's , say $\underline{c}'\underline{y}$, $\underline{c}' = (c_1, \ldots, c_n)$, we call it a linear unbiased estimate. Thus, $\underline{c}'\underline{y}$ is a linear unbiased estimate of $\underline{\ell}'\underline{p}$ if and only if

$$E(\underline{c}'\underline{y}) = \underline{\ell}'\underline{p} \quad .$$

Since $E(\underline{c}'\underline{y}) = \underline{c}'A\,\underline{p}$ we have from the previous equation

$$\underline{c}'A\,\underline{p} = \underline{\ell}'\underline{p}$$

as a necessary and sufficient condition for $\underline{c}'\underline{y}$ to be a linear unbiased estimate of $\underline{\ell}'\underline{p}$. Since we shall consider all of Euclidian m-space as the parameter space, we have equivalently

$$(5) \qquad \underline{c}'A = \underline{\ell}' \quad .$$

We define $V(A')$ to be the vector space generated by the rows of the $m \times n$, $m \leq n$, matrix A' and $V^*(A')$ to the vector space orthogonal to $V(A')$ in the n dimensional vector space over the real numbers.

The following theorem is basic in the study of linear estimation.

THEOREM A: If $\underline{\ell}'\underline{p}$ is a linear combination of the parameters for which there exists a vector $\underline{d}'$ such that $E(\underline{d}'\underline{y}) = \underline{\ell}'\underline{p}$ then there exists exactly one vector $\underline{c}'$ in $V(A')$ for which $E(\underline{c}'\underline{y}) = \underline{\ell}'\underline{p}$. Furthermore, $\operatorname{Var}(\underline{c}'\underline{y})$ minimizes the variance of $\underline{d}'\underline{y}$ over all $\underline{d}'$ such that $E(\underline{d}'\underline{y}) = \underline{\ell}'\underline{p}$.

PROOF: To prove the first assertion consider the decomposition

$$\underline{d}' = \underline{c}' + \underline{e}'$$

where $\underline{c}'$ is the projection of $\underline{d}'$ on $V(A')$ and $\underline{e}'$ the projection of $\underline{d}'$ on $V^*(A')$. Now by assumption

$$\underline{\ell}'\underline{p} = E(\underline{d}'\underline{y})$$

but

$$E(\underline{d}'\underline{y}) = \underline{d}'A\underline{p} = (\underline{c}' + \underline{e}')A\underline{p} = \underline{c}'A\underline{p} + \underline{e}'A\underline{p} = \underline{c}'A\underline{p} = E(\underline{c}'\underline{y})$$

since $\underline{e}'$ is orthogonal to the column vectors of A. Thus,

$$E(\underline{c}'\underline{y}) = E(\underline{d}'\underline{y}) = \underline{\ell}'\underline{p} \quad .$$

Now suppose $\underline{c}_1'$ belongs to $V(A')$ and $E(\underline{c}_1'\underline{y}) = \underline{\ell}'\underline{p}$. Then for every p

$$E(\underline{c}_1'\underline{y}) = E(\underline{c}'\underline{y})$$

or

$$\underline{c}_1' A \underline{p} = \underline{c}' A \underline{p} \qquad \text{for all } p$$

which implies $(\underline{c}_1' - \underline{c}')$ is orthogonal to $V(A')$, i.e., belongs to $V^*(A')$. However, $\underline{c}_1' - \underline{c}'$ belongs to $V(A')$ since each does and thus $\underline{c}_1' - \underline{c}' = \underline{o}'$, i.e., $\underline{c}_1' \equiv \underline{c}'$. This completes the proof of the first assertion.

Now suppose $\underline{d}'\underline{y}$ is a linear unbiased estimate of $\underline{\ell}'\underline{p}$. Then decompose $\underline{d}'$ into $\underline{c}_1'$ and $\underline{e}_1'$ where $\underline{c}_1'$ belongs to $V(A')$ and $\underline{e}_1'$ belongs to $V^*(A')$. As before $\underline{c}_1'\underline{y}$ is also a linear unbiased estimate of $\underline{\ell}'\underline{p}$ and $\underline{c}_1'$ belongs to $V(A')$. By the uniqueness argument given previously $\underline{c}_1' = \underline{c}'$. Hence, $\underline{d}' = \underline{c}' + \underline{e}_1'$. Thus, $\mathrm{Var}(\underline{d}'\underline{y}) = \underline{d}'\sigma^2 I_n \underline{d} = \sigma^2\underline{d}'\underline{d} = \sigma^2(\underline{c}' + \underline{e}_1')(\underline{c} + \underline{e}_1)$ $= \sigma^2\underline{c}'\underline{c} + \sigma^2\underline{e}_1'\underline{e}_1 = \mathrm{Var}(\underline{c}'\underline{y}) + \sigma^2\underline{e}_1'\underline{e}_1$. Therefore $\mathrm{Var}(\underline{d}'\underline{y}) > \mathrm{Var}(\underline{c}'\underline{y})$ for $\underline{d}' \neq \underline{c}'$, i.e., $\underline{e}_1'\underline{e}_1 \neq 0$. This completes the proof.

We shall call this unique estimate which minimizes the variance over all linear unbiased estimates the best estimate of $\underline{\ell}'\underline{p}$.

Theorem A says that if the "best" estimate of $\ell'p$ is $c'y$ then $c' = q'A'$ for some q'. From equation (5) we see that we must have $q'A'A = \ell'$. These equations are called the conjugate normal equations. Conversely, we have that if $q'A'A = \ell'$ then $q'A'y$ is the unique "best" estimate of $\ell'p$.

<u>THEOREM B</u>: (Gauss-Markov) If $\ell'p$ has an unbiased linear estimate then the best estimate is $\ell'\hat{p}$ where $\hat{p}$ are the least squares estimates of p.

<u>PROOF</u>: The least squares estimates of p are those values for p_1, $p_2, \dots, p_n$ which minimize the sum of squared deviations of $y_1, y_2, \dots, y_n$ from their (estimated) expected value. Thus

$$S^2 = \sum_{j=1}^{n} (y_j - a_{j1}p_1 - a_{j2}p_2 - \dots - a_{jm}p_m)^2$$

is to be minimized by choice of $p_1, p_2, \dots p_m$. Now

$S^2 = (y' - p'A')(y - Ap) = y'y - p'A'y - y'Ap + p'A'Ap = y'y - 2p'A'y + p'A'Ap$.

By differentiating S' with respect to each of the p's and setting these m derivatives equal to zero we obtain

$$- 2A'y + 2A'Ap = \underline{0} \quad \text{or}$$

$$* \qquad A'Ap = A'y \quad .$$

Equations * are called the normal equations. Thus, if $\hat{p}$ satisfies the normal equations then $\hat{p}$ is a critical point of S^2 . Now we shall show that it is a minimum point.

Let y' be decomposed as $y' = m' + e'$ where m' is in $V(A')$ and e' is in $V^*(A')$. Thus, $m' = x'A'$ and $e'A = 0$. Then, $y'A = x'A'A + e'A = x'A'A$ or $A'Ax = A'y$. Hence, $\underline{x}$ must satisfy the normal equations. Conversely since $\hat{p}$ satisfies the normal equations, $\hat{p}'A'$ is the projection of y' on $V(A')$ and hence $m' = \hat{p}'A'$. That is $(y' - \hat{p}'A')A = \underline{0}'$ and $\hat{p}'A'$ is in $V(A')$.

<u>COROLLARY</u>: If $Eq'A'y = \ell'p$ then $q'A'y = \ell'\hat{p}$ where $\hat{p}$ are least squares estimates of p .

<u>PROOF</u>: $q'A'$ is in $V(A')$ and by assumption $Eq'A'y = \ell'p$. Hence by theorem A $q'A'y$ is the unique best estimate of $\ell'p$. By theorem B , $\ell'\hat{p}$ is the unique best estimate of $\ell'p$. Hence $q'A'y = \ell'\hat{p}$.

APPENDIX B: PROOF OF THEOREM 1

Consider the rectangular design

$$D = D_1 \times \ldots \times D_\pi \quad , \qquad D_i = \left\{ x_{t_i}^{(i)} \; : \; t_i = 1 \, , \, \ldots \, , \, T_i \right\}$$

and the polynomial (2) written, in preparation for step one of the step procedure, as

$$(3) \qquad b_1^{(\pi)} + b_2^{(\pi)} \, x^{(\pi)} + \ldots + b_d^{(\pi)} \, (x^{(\pi)})^{d-1} + b_{d+1}^{(\pi)} \, (x^{(\pi)})^d$$

where $b_k^{(\pi)}$ is a polynomial in $x^{(1)}, \ldots , x^{(\pi-1)}$ of degree $(d - (k-1))$.

Let $y_{t_1 \ldots t_\pi} = f \, (x_{t_1}^{(1)}, \ldots , x_{t_\pi}^{(\pi)})$ where f is the function to be approximated by the polynomial (2) .

If we can demonstrate that the step procedure estimate of a coefficient is a linear combination of the components of $X'\underline{y}$ --where the matrix X arises from writing the system

$$\begin{aligned} E \, (y_{t_1 \ldots t_\pi}) = \; & a_{11\ldots1} + \ldots + a_{1\ldots i \ldots 1} x_{t_i}^{(i)} + \ldots + \\ & (\text{terms in } x_{t_1}^{(1)}, \ldots , x_{t_\pi}^{(\pi)} \text{ of degree} \leq d) + \\ & a_{11\ldots1d+1} (x_{t_\pi}^{(\pi)})^d \quad , \qquad t_i = 1 \, , \, \ldots \, , \, T_i \end{aligned}$$

in the matrix form

$$E \, (\underline{y}) = X \, \underline{a}$$

as in the case of the preceding example--and show that such an estimate is unbiased; then the step procedure estimate is the least squares estimate [Ref. 1].

Since the first step of the step procedure is a least squares estimation of the $b^{(\pi)}$'s with $x^{(1)}, \ldots, x^{(\pi-1)}$ held fixed, the expected value of the estimate $\hat{b}^{(\pi)}_{d+1}$ of $b^{(\pi)}_{d+1}$ is $b^{(\pi)}_{d+1} = a_{1\ldots 1d+1}$. Also the estimate $\hat{b}^{(\pi)}_{d+1}$ is itself a linear combination of

$$\sum_{t_\pi=1}^{T_\pi} y_{t_1\ldots t_\pi} , \sum_{t_\pi=1}^{T_\pi} x^{(\pi)}_{t_\pi} y_{t_1\ldots t_\pi} , \sum_{t_\pi=1}^{T_\pi} (x^{(\pi)}_{t_\pi})^d \, y_{t_1\ldots t_\pi} .$$

Since $b^{(\pi)}_{d+1}$ is independent of $x^{(1)}, \ldots, x^{(\pi)}$ succeeding steps in the step procedure will at each stage give the mean of the result of the previous stage over the number of data points in the present stage so that the step procedure estimate of $b^{(\pi)}_{d+1}$ is

$$\frac{\sum_{t_1=1}^{T_1} \ldots \sum_{t_{\pi-1}=1}^{T_{\pi-1}} \hat{b}^{(\pi)}_{d+1}}{T_1\, T_2 \, \ldots \, T_{\pi-1}} .$$

Since $\hat{b}^{(\pi)}_{d+1}$ is unbiased, this estimate will be unbiased. This estimate will also be a linear combination of

$$\sum_{t_1=1}^{T_1} \ldots \sum_{t_\pi=1}^{T_\pi} y_{t_1\ldots t_\pi} , \quad \sum_{t_1=1}^{T_1} \ldots \sum_{t_\pi=1}^{T_\pi} x^{(\pi)}_{t_\pi} y_{t_1\ldots t_\pi} , \ldots, \quad \sum_{t_1=1}^{T_1} \ldots \sum_{t_\pi=1}^{T_\pi} (x^{(\pi)}_{t_\pi})^d \, y_{t_1\ldots t_\pi} ,$$

i.e., the components of $X'\underline{y}$. This completes the proof.

It is clear from the proof that by choosing $x^{(i)}$ as the leading variable the step procedure could be used to calculate the least squares estimate of the coefficient of $(x^{(i)})^d$. We are usually interested in the least squares estimate of all the coefficients and in this case theorem 1 is not very helpful.

APPENDIX C: PROOF OF THEOREM 2

Consider the balanced polynomial

$$\sum_{\ell_1=1}^{L_1+1} \cdots \sum_{\ell_\pi=1}^{L_\pi+1} a_{\ell_1 \ldots \ell_\pi} x_{\ell_1}^{(1)} \cdots x_{\ell_\pi}^{(\pi)} \quad , \quad x_{\ell_i}^{(i)} = (x^{(i)})^{\ell_i - 1}$$

as an approximation of a real function f of π real variables $x^{(1)}, \ldots, x^{(\pi)}$ over the rectangular design

$$D = D_1 \times \ldots \times D_\pi \quad , \quad D_i = \left\{ x_{t_i}^{(i)} : t_i = 1, 2, \ldots, T_i \; \middle| \; x_{t_i}^{(i)} \neq x_{s_i}^{(i)} \, , \; t_i \neq s_i \right\} .$$

Let

$$y_{t_1 \ldots t_\pi} = f\,(x_{t_1}^{(1)}, \ldots, x_{t_\pi}^{(\pi)}) \quad .$$

In what follows we shall use capital letters without affixes to denote the appropriate collection of lower case letters for subscripting purposes, e.g.

$$L = \left\{ \ell_1, \ldots, \ell_\pi \right\} .$$

First we shall consider the least squares criterion for estimates of the a's and generate the normal equations, then we shall show that the step procedure estimates of the a's satisfy the normal equations and are thus least squares estimates.

<u>ACTUAL PROOF</u>: Define S by

$$S \triangleq \sum_{t_1 \ldots t_\pi} \left\{ y_{t_1 \ldots t_\pi} - \sum_{\ell_1 \ldots \ell_\pi} a_{\ell_1 \ldots \ell_\pi} x_{t_1 \ell_1}^{(1)} \cdots x_{t_\pi \ell_\pi}^{(\pi)} \right\}^2$$

and calculate $\frac{\partial}{\partial a_{\alpha_1 \ldots \alpha_\pi}} S$. Setting this partial derivative equal to zero, we arrive at the equation

$$\sum_T y_T \, x^{(1)}_{t_1\alpha_1} \cdots x^{(\pi)}_{t_\pi\alpha_\pi} = \sum_L a_L \sum_T x^{(1)}_{t_1\ell_1} x^{(1)}_{t_1\alpha_1} \cdots x^{(\pi)}_{t_\pi\ell_\pi} x^{(\pi)}_{t_\pi\alpha_\pi} .$$

Now employing the properties of the rectangular design we have

$$(4) \quad \sum_T y_T \, x^{(1)}_{t_1\alpha_1} \cdots x^{(\pi)}_{t_\pi\alpha_\pi} = \sum_L a_L \left\{ (\sum_{t_1} x^{(1)}_{t_1\ell_1} x^{(1)}_{t_1\alpha_1}) \cdots (\sum_{t_\pi} x^{(\pi)}_{t_\pi\ell_\pi} x^{(\pi)}_{t_\pi\alpha_\pi}) \right\} .$$

If we let the matrix $(x^{(i)}_{t_i\ell_i})$ be denoted by X_i , then we have that

$$X^{(i)}_{\ell_i\alpha_i} \triangleq \sum_{t_i} x^{(i)}_{t_i\ell_i} x^{(i)}_{t_i\alpha_i}$$

is the element in the ℓ_i , α_i position of the matrix $X_i'X_i$ and from (4)

$$(6) \qquad \sum_T y_T \, x^{(1)}_{t_1\alpha_1} \cdots x^{(\pi)}_{t_\pi\alpha_\pi} = \sum_L a_L \left(X^{(1)}_{\ell_1\alpha_1} \cdots X^{(\pi)}_{\ell_\pi\alpha_\pi} \right) .$$

Equation (6) is a typical equation from the normal equations.

We shall now use the step procedure to estimate a typical a, $a_{s_1 \ldots s_\pi}$. In order to facilitate the writing down of this estimate we shall have need of the following notation. Let

$$(Z^{(i)}_{\ell_i r_i}) \triangleq X_i' \quad , \quad \ell_i = 1 , \ldots , L_i + 1 , \ r_i = 1 , \ldots , T_i$$

$$(X^{\ell_i s_i}_{(i)}) \triangleq (X_i' X_i)^{-1} \quad s_i , \ell_i = 1 , \ldots , L_i + 1 .$$

The first step of the step procedure for finding the estimate of $a_{s_1, \ldots, s_\pi}$ is

$$a^{(1)}_{t_1 \ldots t_{\pi-1} s_\pi} = (X_\pi' X_\pi)^{-1} X_\pi' Y_{t_1 \ldots t_{\pi-1}} = \sum_{r_\pi=1}^{T_\pi} \left\{ \sum_{\ell_\pi=1}^{L_\pi+1} X^{\ell_\pi s_\pi}_{(\pi)} Z^{(\pi)}_{\ell_\pi r_\pi} \right\} y_{t_1 \ldots t_{\pi-1} r_\pi}$$

where $Y'_{t_1\ldots t_{\pi-1}} = (y_{t_1,\ldots,t_{\pi-1},1}\ , \ \ldots\ ,\ y_{t_1,\ldots,t_{\pi-1},T_\pi})$. The second step is then

$$a^{(2)}_{t_1\ldots t_{\pi-2}s_{\pi-1}s_\pi} = (X'_{(\pi-1)}X_{(\pi-1)})^{-1}\ X'_{\pi-1}a^{(1)}_{t_1\ldots t_{\pi-2}s_\pi}$$

$$= \sum_{r_{\pi-1}=1}^{T_{\pi-1}} \left\{ \sum_{\ell_{\pi-1}=1}^{L_{\pi-1}+1} X^{\ell_{\pi-1}s_{\pi-1}}_{(\pi-1)}\ Z^{(\pi-1)}_{\ell_{\pi-1}r_{\pi-1}} \right\} a^{(1)}_{t_1\ldots t_{\pi-2}r_{\pi-1}s_\pi} \quad .$$

The i^{th} step is thus

$$a^{(i)}_{t_1\ldots t_{\pi-i},s_{\pi-i-1},\ldots\ s_\pi} = (X'_{\pi-i-1}X_{\pi-i-1})^{-1}X'_{\pi-i-1}a^{(i-1)}_{t_1\ldots t_{\pi-i},s_{\pi-i-2},\ldots\ s_\pi} \quad .$$

Finally

$$a_{s_1\ldots s_\pi} = (X_1'X_1)^{-1}X'_{(1)}\ a^{(\pi-1)}_{s_2\ldots s_\pi}$$

$$= \sum_{r_1=1}^{T_1} \left\{ \sum_{\ell_1=1}^{L_1+1} X^{\ell_1s_1}_{(1)}\ Z^{(1)}_{\ell_1r_1} \right\} a^{(\pi-1)}_{r_1s_2\ldots s_\pi}$$

$$= \sum_{r_1=1}^{T_1}\ \sum_{\ell_1=1}^{L_1+1} X^{\ell_1s_1}_{(1)}Z^{(1)}_{\ell_1r_1}\ \sum_{r_2=1}^{T_2}\ \sum_{\ell_2=1}^{L_2+1}\ X^{\ell_2s_2}_{(2)}Z^{(2)}_{\ell_2r_2}\quad \cdot\ \cdot\ \cdot$$

$$\sum_{r_{\pi-1}=1}^{T_{\pi-1}}\ \sum_{\ell_{\pi-1}=1}^{L_{\pi-1}+1} X^{\ell_{\pi-1}s_{\pi-1}}_{(\pi-1)}Z^{(T-1)}_{\ell_{\pi-1}r_{\pi-1}}\ \sum_{r_\pi=1}^{T_\pi}\ \sum_{\ell_\pi=1}^{L_\pi+1}\ X^{\ell_\pi s_\pi}_{(\pi)}\ Z^{(\pi)}_{\ell_\pi r_\pi}\ y_{r_1\ldots r_\pi} \quad .$$

By using the fact that we have a balanced polynomial over a rectangular design we can write

$$a_s = a_{s_1\ldots s_\pi} = \sum_R \left\{ y_R \sum_L X^{\ell_1s_1}_{(1)}\ldots\ X^{\ell_\pi s_\pi}_{(\pi)}\ Z^{(1)}_{\ell_1r_1}\ \ldots\ Z^{(\pi)}_{\ell_\pi r_\pi} \right\} \quad .$$

If we substitute a_s for a_L in equation (6), the right hand side of (6) becomes

$$\sum_s \left\{ a_s \, (X^{(1)}_{s_1\alpha_1} \cdots X^{(\pi)}_{s_\pi\alpha_\pi}) \right\} =$$

$$\sum_R y_R \left\{ \sum_L [(Z^{(1)}_{\ell_1 r_1} \cdots Z^{(\pi)}_{\ell_\pi r_\pi})(\sum_{s_1} X^{\ell_1 s_1}_{(1)} X^{(1)}_{s_1\alpha_1}) \cdots (\sum_{s_\pi} X^{\ell_\pi s_\pi}_{(\pi)} X^{(\pi)}_{s_\pi\alpha_\pi})] \right\} .$$

However

$$\sum_{s_i} (X^{\ell_i s_i}_{(i)} X^{(i)}_{s_i\alpha_i}) = \delta_{\ell_i\alpha_i} = \begin{cases} 0 \, , & \ell_i \neq \alpha_i \\ 1 \, , & \ell_i = \alpha_i \end{cases} ,$$ so that we have the

right hand side of (6) equal to

$$\sum_R y_R \left\{ \sum_L (Z^{(1)}_{\ell_1 r_1} \cdots Z^{(\pi)}_{\ell_\pi r_\pi}) \; \delta_{\ell_1\alpha_1} \cdots \delta_{\ell_\pi\alpha_\pi} \right\} = \sum_R y_R \, (Z^{(1)}_{\alpha_1 r_1} \cdots Z^{(\pi)}_{\alpha_\pi r_\pi})$$

which is the left hand side of equation (6) . Thus a_s is a solution to the normal equations and the proof is therefore complete.

REFERENCES

[1] Bose, R. C., Least Squares Aspects of Analysis of Variance, University of North Carolina, Institute of Statistics Mimeograph Series No. 9, Chapel Hill, North Carolina.

[2] Cramer, H., Mathematical Methods of Statistics, Princeton University Press (1946).

[3] Loève, M., Probability Theory, D. Van Nostrand Company, Inc. (1955).

N63-20963 267

MINNEAPOLIS-HONEYWELL REGULATOR COMPANY
MILITARY PRODUCTS GROUP
RESEARCH DEPARTMENT

A LYAPUNOV TECHNIQUE FOR OBTAINING EXTENSIONS OF AN OPEN LOOP CONTROL TO A NEIGHBORHOOD OF THE OPEN LOOP TRAJECTORY

By

D. L. Lukes

ST. PAUL, MINNESOTA

A LYAPUNOV TECHNIQUE FOR OBTAINING EXTENSIONS OF AN OPEN LOOP CONTROL TO A NEIGHBORHOOD OF THE OPEN LOOP TRAJECTORY

By D. L. Lukes

Introduction

We assume that a reference trajectory and the corresponding control function have been determined for a given dynamical system. The problem posed is the extension of the control to a neighborhood of the reference trajectory to obtain a feedback control which drives the system to the given final state. The technique is based on the construction of a Lyapunov function defined in some neighborhood of the reference trajectory.

The usual treatment of this problem is to linearize the system equations with respect to the reference path and then control the perturbations about the trajectory. That technique has the advantage that the system of linear perturbations can be readily analyzed, but for nonlinear systems it is usually difficult to provide any simple appraisal of the stability of the over-all procedure. The Lyapunov technique presented differs from the classical technique by not requiring a linearization of the system equations. Further, stability is insured. It is tacitly assumed in both techniques that the over-all design of the system is based upon some nominal trajectory and in order to maintain its validity the control must keep the output of the system in some neighborhood of the open loop (reference) trajectory.

In this preliminary investigation, the approach taken is to look for the characteristic properties of controls which provide the extensions and leave open specific determinations for other system requirements.

The System and the Control Problem

We consider dynamical systems of the form

$$\frac{dx}{dt} = g(x, u) ,$$

where x and g are finite dimensional vectors and u is the control vector to be determined. It will always be assumed that the systems are autonomous (t does not occur explicitly in g).

Furthermore, to simplify the exposition, we will take x to be two-dimensional and u to be one-dimensional. Thus, we will consider a system

$$\frac{dx_1}{dt} = g_1(x, u)$$

$$\frac{dx_2}{dt} = g_2(x, u) .$$

It will be advantageous to use arc length as the independent variable rather than time, so we define

$$s(t) = \int_0^t |g| \, d\tau ,$$

assuming that $g \neq 0$, where $|g| = \sqrt{g_1^2 + g_2^2}$. We then get the system

$$\frac{dx_1}{ds} = f_1(x, u) \qquad\qquad \left(f = \frac{g}{|g|}\right)$$

$$\frac{dx_2}{ds} = f_2(x, u) \quad .$$

Without loss of generality we assume that $x(0) = 0$.

Let the given reference control function be represented by $u^\circ = u^\circ(x^\circ(s))$ and the open loop trajectory by $x^\circ = x^\circ(s)$ for $0 \leq s \leq L$. (L is the total length of the open loop trajectory.)

The problem is to extend u° to $u = u(x)$ on some neighborhood of the reference so that

(a) $u(x^\circ) = u^\circ(x^\circ)$

(b) with the control $u = u(x)$ the trajectory remains in the neighborhood of x° and passes through $x(L)$, as illustrated in figure 1.

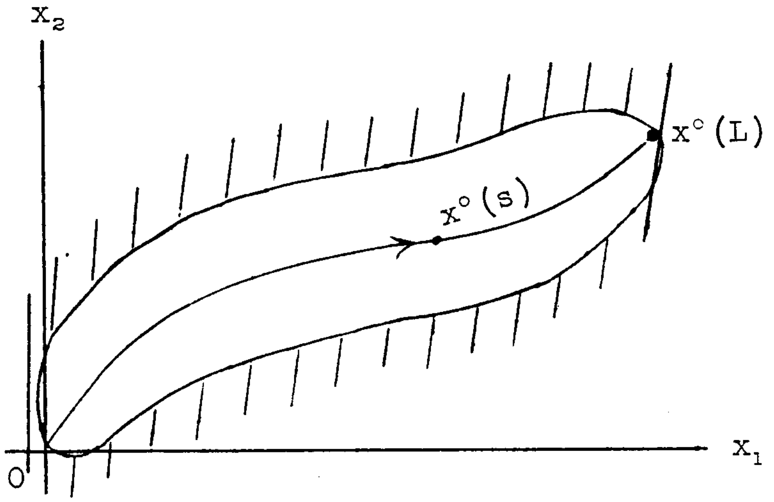


FIGURE 1: A NEIGHBORHOOD OF THE REFERENCE TRAJECTORY

Coordinates for the Neighborhood of the Reference Trajectory

In order to assign coordinates we make the standard definitions used in the differential geometry of space curves. Let the unit tangent be designated by

$$\hat{T}(s) = f(x^\circ, u^\circ) \quad ,$$

the unit normal by

$$\hat{N}(s) = \frac{\hat{T}'(s)}{|\hat{T}'(s)|} \qquad (\text{where } k \neq 0)$$

and the curvature symbolized by

$$k(s) = |\hat{T}'(s)| \quad .$$

Then $\hat{T}(s)$ is a unit vector tangent to the reference path at the point $x^\circ(s)$ and $\hat{N}(s)$ is orthogonal to $\hat{T}$ at $x^\circ(s)$. The curvature $k(s)$ is a measure of how $\hat{T}(s)$ changes its direction as we move along x°.

Now be letting $y_1 = 2s$ and setting

$$x = Ry_2\hat{N}+x^\circ \quad ,$$

where R is a fixed positive constant we can assign the new coordinates (y_1 , y_2) to every point in a neighborhood of x° (see figure 2).

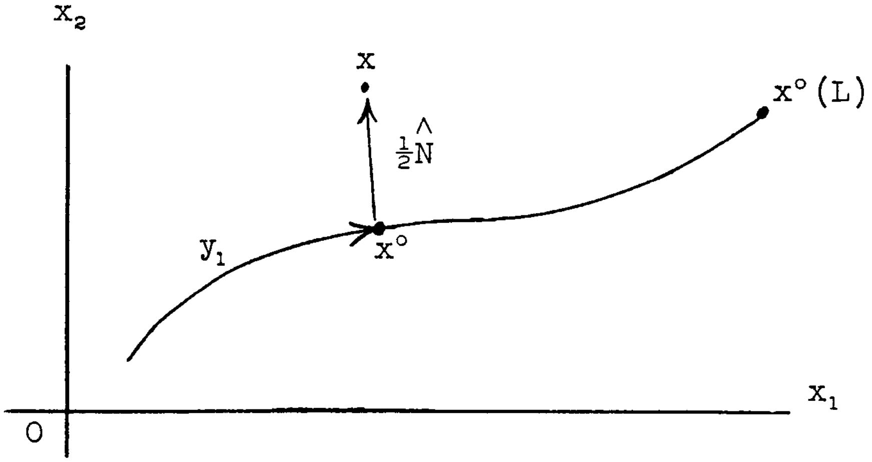


Figure 2: Coordinates in a Neighborhood of the Reference Trajectory

If x° is simple and differentiable then the mapping will be a homomorphism on some neighborhood of x°. With this mapping we can stretch an elliptical region along the trajectory as indicated in figure 3.

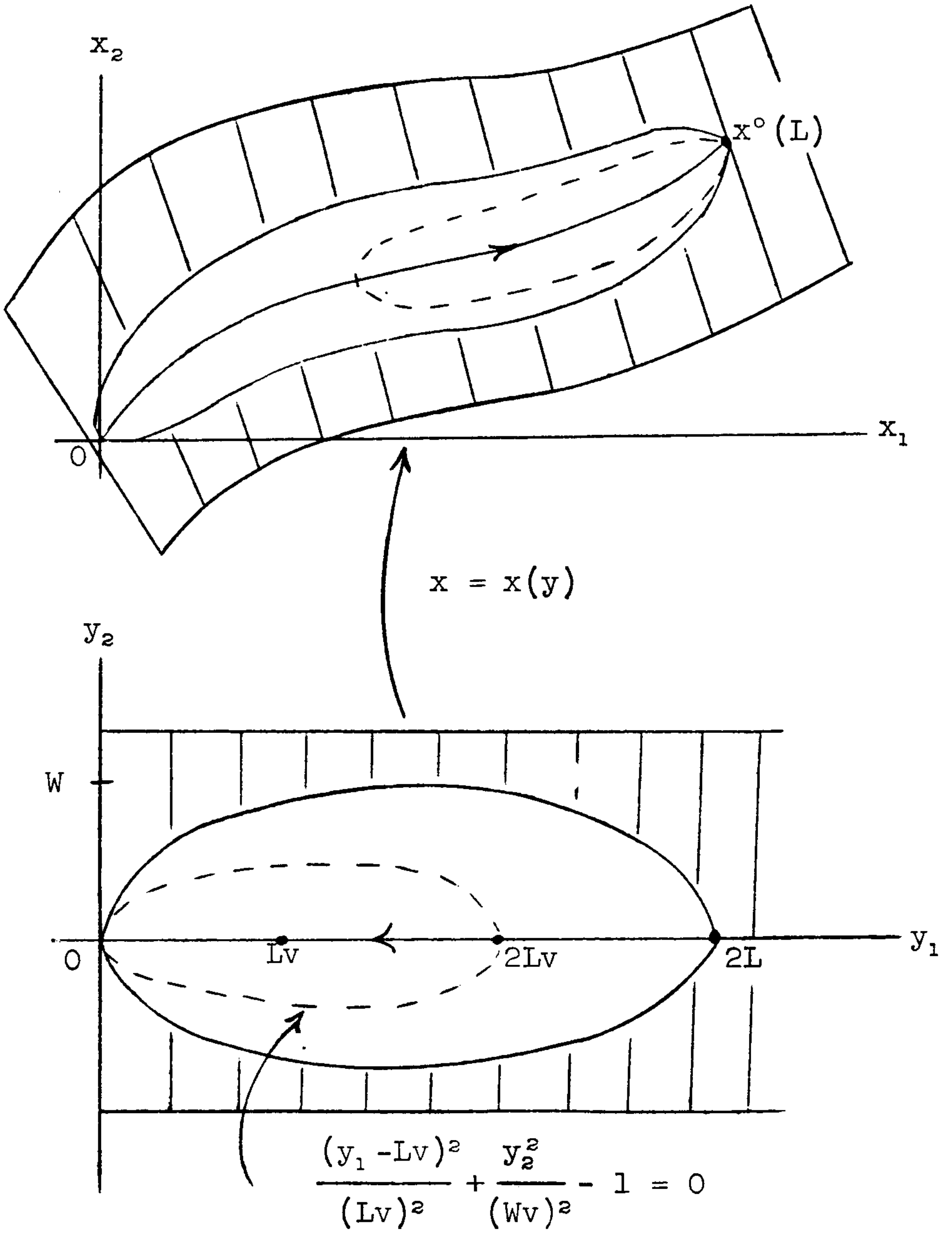

x_2
$x^\circ(L)$
x_1
O
$x = x(y)$
y_2
W
O
Lv
2Lv
2L
y_1
$\frac{(y_1 - Lv)^2}{(Lv)^2} + \frac{y_2^2}{(Wv)^2} - 1 = 0$

FIGURE 3: MAPPING AN ELLIPTICAL NEIGHBORHOOD ALONG x°

Then by writing the equation for a one parameter family of retracting ellipses in the y-plane, we obtain a Lyapunov function on the neighborhood of control in the x-plane (see figure 3).

Lyapunov Function

A short calculation yields the following expression

$$v(y(x)) = \tfrac{1}{2}\left[\left(\frac{y_1}{L}\right) + \left(\frac{y_2}{W}\right)^2 \left(\frac{L}{y_1}\right)\right]$$

which serves as a Lyapunov function. The coordinates are related by the formula

$$x(y) = Ry_2\hat{N}(y_1) + x^\circ(y_1) \ .$$

On a neighborhood of x° we can solve this equation to get $y = y(x)$.

The retracting neighborhood of x° in the y-plane can be described by the inequalities

$$0 \leq y_1 \leq 2Lv$$

$$y_2 \leq W\sqrt{\left(\frac{y_1}{L}\right)\left(2v - \frac{y_1}{L}\right)} \ .$$

Now to get a relation between u and v which guarantees that the resulting trajectory will fall within the neighborhood of control and passes through $x^\circ(L)$, we calculate $dv(y)/ds$ along a trajectory $x(s)$. Differentiating the formula

$$v = \tfrac{1}{2}\left[\left(\frac{y_1}{L}\right) + \left(\frac{y_2}{W}\right)^2 \left(\frac{L}{y_1}\right)\right]$$

yields

$$\frac{\partial v}{\partial y_1} = \tfrac{1}{2}\left[\left(\frac{1}{L}\right) - \left(\frac{y_2}{W}\right)^2 \left(\frac{L}{y_1^2}\right)\right] \ .$$

Further

$$\frac{\partial v}{\partial y_2} = \left(\frac{y_2}{W^2}\right)\left(\frac{L}{y_1}\right) .$$

and since

$$\frac{dv}{ds} = \left[\frac{\partial v}{\partial y_1} + \frac{\partial v}{\partial y_2}\,\frac{dy_2}{dy_1}\right]\left(\frac{dy_1}{ds}\right)$$

$$= \left[\left(\frac{1}{L}\right) - \left(\frac{y_2}{W}\right)^2\left(\frac{L}{y_1^2}\right)\right] + 2\left(\frac{y_2}{W^2}\right)\left(\frac{L}{y_1}\right)\frac{dy_2}{dy_1} ,$$

we find

$$\left(\frac{dv}{ds} - \frac{1}{L}\right) = 2\left(\frac{y_2}{W^2}\right)\left(\frac{L}{y_1}\right)\left[\frac{dy_2}{dy_1} - \frac{1}{2}\,\frac{y_2}{y_1}\right]$$

$$= \left(\frac{y_2}{W^2}\right)\left(\frac{L}{y_1}\right)\left[f(x,\ u) \cdot \nabla_X y_2 - \frac{y_2}{y_1}\right] ,$$

where

$$\nabla_X y_2 = \left(\frac{\partial y_2}{\partial x_1} , \frac{\partial y_2}{\partial x_1}\right) .$$

Note that along x°, $y_2 = 0$ so $\frac{dv}{ds} = \frac{1}{L} > 0$. This corresponds to the fact that $x^\circ(s)$ passes through $x^\circ(L)$ as s goes to L.

We know that as long as $\frac{dv}{ds} > 0$ and v is a continuously differentiable function of x, the trajectory will pass through $x^\circ(L)$. Thus our control criterion is

$$\left(\frac{y_2}{W^2}\right)\left(\frac{L}{y_1}\right)\left[f(x,\ u) \cdot \nabla_X y_2 - \frac{y_2}{y_1}\right] + \frac{1}{L} > 0 .$$

Since y is expressible in terms of x, this is an inequality on x and u that must be satisfied in a neighborhood of x°. It should be noted that $\nabla_X y_2$ can be calculated in terms of $\hat{T}$, $\hat{N}$ and k by using the Frenet formulas:

$$\hat{T}' = k\hat{N}$$

$$\hat{B}' = -\tau\hat{N}$$

$$\hat{N}' = -k\hat{T} + \tau\hat{B} \quad ,$$

where $\hat{B}$ is the unit binormal and τ is the torision. Of course, since we are dealing with curves in the plane only the first is needed.

It should be noted that one particular choice of the control which satisfies the above inequality makes $f(x,\ u) \cdot \nabla_X y_2 - \frac{y_2}{y_1} = 0$. This is the same as requiring

$$\frac{dy_2}{dy_1} - \frac{1}{2}\frac{y_2}{y_1} = 0.$$

Solving this differential equation yields $y_1 = \frac{y_2^2}{c^2}$, where c is a constant determined by the deviation of the trajectory from the reference path. In this case the control satisfies the equation

$$f(x, u) \cdot \nabla_x y_2 - \frac{c^2}{y_2} = 0$$

Thus, the control is calculated in terms of only one of the y-coordinates.

When the state of the system x coincides with the reference trajectory, we can define u as u°. This may result in a discontinuous control function. An alternative is to abandon the original reference control after the feedback control has been calculated.

This is as far as the technique can be carried for a general dynamical system. When a specific system is chosen, the proper solutions of the control inequality will become apparent.

Summary

The construction of a Lyapunov function in some neighborhood of a reference trajectory using natural coordinates has been illustrated. With this function it is possible to specify a sufficiency condition in the form of an inequality in terms of the control function u and the state of the system x. This inequality isolates classes of controls which will drive the system to the end state as specified by the reference control.

Furthermore, these controls can be made to coincide with the reference control along the reference trajectory if discontinuous control functions are allowed.

The technique does not require the linearization of the system equations and therefore provides an alternative method for determining controls for some neighborhood of a design trajectory.

Finally, it should be noted that the same technique can be applied to an n-dimensional system with a vector control function.

Example

Consider the dynamical system represented by the equations

$$\frac{dx_1}{dt} = x_2 - 1$$

$$\frac{dx_2}{dt} = 1 - x_1 + u$$

with a reference control function $u^\circ = 2(x_1^\circ - 1)$. In terms of arc length, the equations are

$$\frac{dx_1}{ds} = \frac{x_2 - 1}{\sqrt{(x_2 - 1)^2 + (1 - x_1 + u)^2}}$$

$$\frac{dx_2}{ds} = \frac{1 - x_1 + u}{\sqrt{(x_2 - 1)^2 + (1 - x_1 + u)^2}}$$

so the reference trajectory is

$$x^\circ(s) = \frac{s}{\sqrt{2}}\begin{pmatrix}1\\1\end{pmatrix} \quad 0 \leq s \leq \sqrt{2} = L .$$

We find the tangent and normal vectors to be

$$\hat{T}(s) = \frac{1}{\sqrt{2}}\begin{pmatrix}1\\1\end{pmatrix}, \quad \hat{N}(s) = \frac{1}{\sqrt{2}}\begin{pmatrix}-1\\1\end{pmatrix} \quad \text{and} \quad k(s) = 0.$$

Thus, the coordinate transformation used is

$$x(y) = Ry_2\hat{N} + x^\circ$$

so

$$x(y) = \begin{pmatrix}\frac{1}{2\sqrt{2}} & \frac{-R}{\sqrt{2}}\\ \frac{1}{2\sqrt{2}} & \frac{R}{\sqrt{2}}\end{pmatrix}\begin{pmatrix}y_1\\ y_2\end{pmatrix}$$

and the inverse is

$$y(x) = \begin{pmatrix} \frac{2}{\sqrt{2}} & \frac{2}{\sqrt{2}} \\ \frac{-1}{\sqrt{2}\ R} & \frac{1}{\sqrt{2}\ R} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

Thus, $\nabla_X y_2 = \begin{pmatrix} \frac{-1}{\sqrt{2}\ R} \\ \frac{1}{\sqrt{2}\ R} \end{pmatrix}$ and we can substitute into the control inequality

$$\left(\frac{y_2}{W^2}\right)\left(\frac{L}{y_1}\right)\left[f(x,\ u) \cdot \nabla_X y_2 - \frac{y_2}{y_1}\right] + \frac{1}{L} > 0 \quad .$$

A particular solution is obtained by setting

$f(x,\ u) \cdot \nabla_X y_2 - \frac{y_2}{y_1} = 0.$ This yields the equation

$$\frac{2 - x_1 - x_2 + u}{\sqrt{(x_2 - 1)^2 + (1 - x_1 + u)^2}} = \sqrt{2}\left(\frac{x_2 - x_1}{x_2 + x_1}\right)$$

By solving a quadratic equation, a solution $u = u(x)$ on some neighborhood of the reference trajectory is obtained. Notice that on the reference path the solution agrees with the reference control.

MARTIN MARIETTA COMPANY
Denver, Colorado

SR 0520-54

CONTROLLABILITY FOR LINEAR AND NONLINEAR SYSTEMS

Prepared by:

Dr. H. Hermes
Applied Mathematics Section

Approved by:

R S Novosad

Dr. R. S. Novosad
Chief
Applied Mathematics Section
Electronics Laboratory

September 1963

Section Reports are informal ready records of studies and data developed in connection with the work of the Electronics Laboratory. They may be modified or withdrawn when better data become available. Section Reports contain the author's views only and are not official policy papers of the Martin Company or of the Air Force.

CONTROLLABILITY FOR LINEAR AND NONLINEAR SYSTEMS

H. Hermes [+]

Introduction

The concept of complete controllability was introduced by R. E. Kalman for linear systems. It is the purpose of this work to give a method of extending this notion to nonlinear systems with control appearing linearly.

The motivation for the method of extension came largely from results obtained in [3], and from the geometric interpretations of non-integrability of pfaffians given in [1] and [2]. In particular, Caratheodory gives an argument to show that if, for a single pfaffian equation, there are points in every neighborhood of a given point which are not "reachable" from the given point by curves satisfying the equation, the equation is integrable. This was generalized to systems of pfaffians in [2]. There is a difficulty in trying to apply such results to pfaffian systems which are quite naturally associated with control systems having linear control. (See § II). The reason for this is that the independent variable t, or time, appears explicitly in the pfaffian associated with a control system. Hence the integral curves of the pfaffian system, which can be related back to solutions of the control system, and are used to connect neighboring points to a given point, must have t parametrized as $t(\sigma)$ with $t(\sigma)$ monotone. This is not the case in the proofs given in [1] and [2].

[+] This work was supported in part by NASA Contract No. NAS8-11020.

The relation between singular problems and controllability also arises quite naturally, as can be anticipated from results obtained by LaSalle in [6] . There it is shown that if a linear system is normal, the bang-bang control is unique for the time optimal problem.

Since for a single component of control, normal and proper are equivalent, and proper is equivalent to complete controllability, one expects that complete controllability has a relation to the presence of singular arcs. This is discussed in § II.

I. Complete Controllability for Linear Systems

In this section we will be primarily concerned with the linear time varying system

$$(1.1) \qquad \dot{x}(t) = A(t)x(t) + G(t)u(t)$$

and the constant system

$$(1.2) \qquad \dot{x}(t) = A\,x(t) + G\,u(t)$$

where we assume A(t) is a continuous nxn matrix valued function of t; G(t) is a continuous nxr matrix valued function of t, with $1 \leq r < n$, while the control vector u is a measurable, finite valued, r vector function of t.

The definition of complete controllability for systems of the form (1.1) and (1.2) was given by Kalman, and its consequences have been studied in a series of papers, see in particular [4] , [5] . For the purpose of completeness, I will summarize some of the main results in this section.

Let $\varphi^{u}(t;\ t_o,\ x_o)$ denote the value of the solution of (1.1) at time t, for control vector u, and initial data x_o given at t_o. $\Phi(t, \tau)$ will denote a fundamental solution of the homogeneous equation $\dot{x} = A(t)x$. The same notation will be used when we have system (1.2) in mind.

<u>Definition</u> (Kalman). The system (1.1) is completely controllable at t_o if for every initial state x_o, there is a control u depending on t_o and x_o, such that $\varphi^{u}(t_1;\ t_o,\ x_o) = 0$ for some finite t_1.

This definition is valid for the system (1.2) but reference to t_o is no longer needed.

<u>Theorem 1.1</u> (Kalman). The system (1.1) is completely controllable at time t_o iff the matrix

$$W(t_o,\ t_1) \equiv \int_{t_o}^{t_1} \Phi(t_o,\ \tau)G(\tau)\ G^{T}(\tau)\ \Phi^{T}(t_o,\ \tau)d\tau$$

is non-singular for some $t_1 > t_o$. (Note: W is a symmetric positive semi-definite matrix.)

<u>Proof</u> a) To show sufficiency, assume $W^{-1}(t_o,\ t_1)$ exists. Set $u(\tau) = -G^{T}(\tau)\ \Phi^{T}(t_o, \tau)W^{-1}(t_o,\ t_1)x_o$, for arbitrary initial data x_o. Then

$$\varphi^{u}(t_1;t_o,x_o) = \Phi(t_1,t_o)x_o - \Phi(t_1,t_o)W(t_o,t_1)W^{-1}(t_o,t_1)x_o = 0.$$

<u>Remark 1</u> The condition $W(t_o,t_1)$ non-singular yields a stronger result than required by the definition, i.e., any point x_o can be controlled to the origin in time t_1, where t_1 is independent of x_o.

b) To show necessity, assume the system is completely controllable at t_o. Let $e_1, \ldots, e_n$ be a basis for E^n, and $t_1, \ldots, t_n$ be the corresponding times it takes to control the basis elements to the origin. Let $\bar{t} = \max \{t_1, \ldots, t_n\}$. It will next be shown that any initial value x_o can be controlled to the origin in time $\bar{t}$. Let $u^1, \ldots, u^n$ be the controls which take the basis elements to the origin and define $\bar{u}^1, \ldots, \bar{u}^n$ by

$$\bar{u}^j(t) = \begin{cases} u^j(t) \ ; & t_o \leq t \leq t_j \\ 0 \ ; & t_j < t \leq \bar{t} \ . \end{cases}$$

Now $x_o = \sum \alpha_\nu e_\nu$ for some set of scalars α_ν. Define

$u(t) = \sum \alpha_\nu \bar{u}^\nu(t)$. We will show u takes x_o to the origin in time $\bar{t}$. Indeed

$$\Phi(\bar{t}, t_o) \sum \alpha_\nu e_\nu + \int_{t_o}^{\bar{t}} \Phi(\bar{t}, \tau) G(\tau) \left[\sum \alpha_\nu \bar{u}^\nu(\tau) \right] d\tau =$$

$$\sum \alpha_\nu \left\{ \Phi(\bar{t}, t_o) e_\nu + \int_{t_o}^{\bar{t}} \Phi(\bar{t}, \tau) G(\tau) \bar{u}^\nu(\tau) d\tau \right\} = 0.$$

We now assume $W(t_o, t)$ is singular for all $t > t_o$ and show a contradiction. This assumption implies there exists $x_o \neq 0$ such that $x_o^T W(t_o, \bar{t}) x_o = 0$. Define

$$u^*(t) = -G^T(\tau) \Phi^T(t_o, t) x_o \ .$$

Then

$$x_o^T W(t_o,\bar{t})x_o = \int_{t_o}^{\bar{t}} u*^T(\tau)\ u*(\tau)\,d\tau = 0\ ,$$

which implies $u*(t) \equiv 0$ since $u*$ is continuous. On the other hand, there exists a $\hat{u}$ such that

$$x_o = -\int_{t_o}^{\bar{t}} \Phi(t_o,\tau)G(\tau)\ \hat{u}(\tau)d\tau$$

therefore

$$x_o^T\ x_o = -\int_{t_o}^{\bar{t}} \hat{u}^T(\tau)\ G^T(\tau)\ \Phi^T(t_o,\tau)x_o\ d\tau$$

$$= -\int_{t_o}^{\bar{t}} \hat{u}^T(\tau)\ u*(\tau)d\tau\ = 0$$

since $u* \equiv 0$.

This contradicts the fact that $x_o \neq 0$.

Remark 2. It is easy to show that if the system (1.1) is completely controllable at t_o, then it is completely controllable at any time $t < t_o$. It is not, however, necessarily completely controllable at a time $t > t_o$.

Corollary 1.1 A necessary and sufficient condition that there exist an $(n \times r)$ matrix valued measurable function $V(t)$ such that for some $t_2 > t_o$, the matrix

$$(1\text{-}3) \qquad \overline{W}(t_o, t_2) \equiv \int_{t_o}^{t_2} \Phi(t_o, \tau) G(\tau) V(\tau) d\tau$$

is non-singular, is that for some $t_1 > t_o$, $W(t_o, t_1)$ is non-singular.

Proof: Sufficiency is immediate by choosing

$$V(t) = G^T(t) \Phi^T(t_o, t)$$

and

$$t_1 = t_2.$$

To show necessity, assume $V(t)$ is such that $\overline{W}(t_o, t_2)$ is non-singular. We proceed to show that this implies (1.1) is completely controllable at t_o. Indeed the control

$$u(t) = -V(t) \overline{W}^{-1}(t_o, t_2) \Phi(t_2, t_o) x_o$$

takes the arbitrary initial data x_o, given at t_o, to the origin in time t_2. This implies (1.1) is completely controllable, which, by theorem 1.1 implies there exists a $t_1 > t_o$ such that $W(t_o, t_1)$ is non-singular. ▌

Corollary 1.2 (Kalman) The system (1.2) is completely controllable iff the rank of the matrix $\left[G, AG, \ldots, A^{n-1} G\right] = n$. In this case any point can be controlled to the origin in an arbitrarily small positive interval of time.

Proof: See [4] .

Remark 3. As shown in theorem 1.1, if $W(t_o, t_1)$ is non-singular then any point x_o can be controlled to the origin in time t_1. Also, if $\bar{x}$ is any other point, then $\bar{x}$ can be attained at time t_1, from the point x_o at t_o, by use of the control

$$u(t) = G^T(t)\, \Phi^T(t_o, t)\, W^{-1}(t_o, t_1) \left[\Phi(t_o, t_1)\, \bar{x} - x_o \right].$$

Application to Minimum Amplitude Transfer

Assume that the system

$$(1\text{-}4) \qquad \dot{x}(t) = A(t)\, x(t) + h(t)\, u(t)$$

is completely controllable at t_o, with $W(t_o, t_1)$ non-singular and $\Phi(t, t_o)$ the fundamental solution of $\dot{x} = A(t)\, x$.

The problem considered is that of transferring an arbitrary point x_o to the origin in a given time t_1 (which is large enough so that the transfer is possible) and to do this with a control which has minimum $\mathcal{L}^\infty [t_o, t_1]$ norm. The problems of such transfers with minumum energy, i.e., controls which have minumum $\mathcal{L}^2$ norm is solved in [5] .

It will turn out that the control with minimum $\mathcal{L}^\infty$ norm for the above problem, will be constant, in absolute value, for almost all t, i.e., a bang-bang control. This should be expected, in view of the results obtained by LaSalle [6] for the time optimal problem.

Define $F^i(t) \equiv \sum_{j=1}^{n} \Phi^{ij}(t_o, t)\, h^j(t), \qquad i = 1, 2, \ldots, n,$

and let F be the vector with components F^i. Then the assumption of complete controllability, and t_1 such that $W(t_o, t_1)$ is non-singular, implies that for any constant vector x_o, there exists a control u such that

$$(1\text{-}5) \qquad x_o = \int_{t_o}^{t_1} F(\tau)u(\tau)d\tau \; .$$

We consider the functions F^i in $\mathcal{L}^1\left[t_o, t_1\right]$ and the control u in $\mathcal{L}^\infty\left[t_o, t_1\right]$. Let L be the linear subspace spanned by the functions $F^1, \ldots, F^n$. Define $L^\perp$ by

$$L^\perp = \left\{ g \in L^\infty : \int_{t_o}^{t_1} g(\tau)\, F^i(\tau)\, d\tau = 0,\ i = 1,2,\ldots,n \right\}$$

Let v be any control satisfying (1-5). As the solution we seek a control u, of smallest $\mathcal{L}^\infty$ norm and such that $(u-v) \in L^\perp$, i.e., we seek a closest element $\omega \in L^\perp$ to v, and then set $u = v - \omega$.

The problem is now posed so that the following well known theorem of functional analysis can be applied.

Theorem: Let L be a linear subspace of a normed linear space X and let $L^\perp \subset X^*$ (the normed conjugate space) be the set of continuous linear functionals in X* vanishing on L. For any $x_o^* \in X^*$, of distance d from $L^\perp$, we have

$$d = \min_{\ell^* \in L^\perp} \left| x_o^* - \ell^* \right| = \sup_{x \in L} \frac{|x_o^*\, x|}{|x|} = \left| x_o^* \right|_L$$

where the minimum on the left is actually attained by some ℓ_o^* in $L^\perp$.

(Here $|x_o^*|_L$ denotes the norm of x_o^* on the subspace L.).

For a proof of this theorem, see [8] , where a moment problem of the form (1-5) is also considered, and it is shown that the solution satisfies $|u(t)|$ = const. for almost all t.

II. Extension of Complete Controllability to Non-linear Systems, with Linear Control

In this section we consider extending the notion of complete controllability to systems of the form

$$(2\text{-}1) \qquad \dot{x}(t) = g(t, x(t)) + H(t, x(t))\, u(t)$$

where g is an n-vector, H an $n \times r$ matrix, while u is a finite valued measurable control vector. It is assumed that g and H are C^1 in all arguments. Throughout this section the stipulation $1 \leq r < n$ is required to hold, and it is assumed, mainly for convenience of notation, that H has constant rank r throughout the domain $\mathcal{D}$ in (t, x) space of interest.

Let $B(t, x)$ be a C^1, $(n-r) \times n$, matrix, with rank $n-r$, satisfying

$$(2\text{-}2) \qquad B(t, x)\, H(t, x) \equiv 0, \qquad\qquad (t, x) \epsilon\, \mathcal{D} .$$

We can therefore associate, with a system of the form (2-1), a pfaffian system

$$(2\text{-}3) \qquad B(t, x) dx - B(t, x)\, g(t, x)\, dt = 0 .$$

Definition 2.1. The pfaffian system (2-3) is integrable if there exists a linear combination of the rows of B, taken with C^1 scalar valued coefficients $\alpha_\nu (t, x)$, such that if $b(t, x) = \sum_{\nu=1}^{n-r} \alpha_\nu (t,x) b^\nu (t,x)$, where the b^ν are the rows of B, the pfaffian

$$(2\text{-}4) \qquad b(t,x)dx - b(t,x)g(t,x)dt = 0$$

is an exact differential. (We assume $b \not\equiv 0$.)

It should be noted that there is no loss of generality in assuming (2-4) an exact differential, since if it were merely integrable, the integrating factor could be included with the scaler multipliers α_ν to form a new pfaffian which is an exact differential. Throughout this section the vector b will represent some linear combination of the rows of B.

Before stating an explicity criterion for complete controllability of a system of the form (2-1), one may ask: What should one expect the definition to yield? This can presently be answered as follows. Since the definition should extend that given for the linear systems considered in Section I, which are special cases of (2-1), one expects:

a) If $g(t, x) \equiv A x$, $H(t, x) \equiv G$, where A and G are constant, then the analytic criterion which defines complete controllability for (2-1) should imply and be implied by the rank of the matrix $\left[G, AG, \ldots, A^{n-1} G\right] = n$.

b) If $g(t, x) = A(t)x$, $H(t, x) = H(t)$, then the criterion which defines complete controllability of (2-1) should be equivalent with the condition $W(t_o, t_1)$ is non-singular for some $t_1 > t_o$.

c) There should be a geometric interpretation of the condition, e.g., what points are attainable from the initial point in finite time. In the linear system, there were global attainability results, i.e., any point oould be attained from the initial point via a trajectory of the system. In the non-linear problem, one would expect at most local results of this nature.

The approach will be to state a criterion for complete controllability of (2-1) which we will show satisfies a) and b) above. We then use this criterion to show what the geometric interpretation mentioned in c) should be. Of course, how the definition of complete controllability should be extended, is a matter of personal opinion.

Definition 2.2. The system (2-1) is completely controllable at t_o if the associated pfaffian system (2-2) is not integrable for $t \geq t_o$, $(t, x) \in \mathcal{A}$.

It will next be shown that this criterion is equivalent to the condition $W(t_o, t_1)$ being non-singular for some $t_1 > t_o$, when $g(t, x) = A(t)x$, $H(t, x) = H(t)$.

For the system

(2-5) $$\dot{x} = A(t)x + H(t)u$$

to form the associated pfaffian system, it suffices to take $B = B(t)$. Also, in forming the vector $b = b(t)$, there is no loss of generality in taking the functions α_ν as functions of only t. Indeed we must only show that if the pfaffian

(2-6) $$b(t)dx - b(t)\,A(t)x\,dt = 0$$

is integrable, then the integrating factor, denoted by μ, can be taken as a function of only t. To obtain this, suppose $\bar{\mu}(t, x)$ is such that

$$\bar{\mu}(t, x)\,b(t)\,dx - \bar{\mu}(t, x)\,b(t)\,A(t)x\,dt$$

is an exact differential. Then $\bar{\mu}_{x_j} b^i - \bar{\mu}_{x_i} b^j = 0$ for all $i,j=1,2,..,n$

and $\bar{\mu}_t b + \bar{\mu}\dot{b} = -\bar{\mu}_x bAx - \bar{\mu} bA$. Define $\mu(t) = \bar{\mu}(t, 0)$.

It follows that $\mu(t)$ is also an integrating factor.

Since it is sufficient to consider both μ and the α_ν as functions of only t, there is no loss of generality in considering that if the pfaffian system

(2-7) $\quad B(t)dx - B(t)A(t)x\,dt = 0$

associated with (2-5) is integrable, then (2-6) is an exact differential for some b.

<u>Theorem 2.1</u> For the system (2-5), a necessary and sufficient condition for $W(t_o, t_1)$ to be non-singular for some $t_1 > t_o$, is that the associated pfaffian system (2-7) be non-integrable.

<u>Proof:</u> a) Necessity. (We shall prove the contrapositive).

Assume (2-7) is integrable. This implies (2-6) is an exact differential for some vector b, which in turn implies

$$\dot{b}(t) \equiv -b(t)\ A(t)\ .$$

Let $\Phi(t, t_o)$ be a fundamental solution of $\dot{x} = A(t)\,x$. Then the vector b admits the representation $b(t) = c\,\Phi^{-1}(t, t_o)$ for some constant vector c. Let h(t) be any column of H(t). Then $0 \equiv b(t)h(t) = c\,\Phi^{-1}(t,t_o)h(t) = c\,\Phi(t_o, t)h(t)$. But

$$W(t_o,t_1) = \int_{t_o}^{t_1} \Phi(t_o,t)H(t)H^T(t)\Phi^T(t_o,t)dt.$$

Since h was an arbitrary column of H, for every $t_1 \geq t_o$ we have

$c\, W(t_o, t_1)\, c^T \equiv 0$ which implies since W is a symmetric, positive semi definite matrix, that W is singular for every $t_1 \geq t_o$.

b) Sufficiency (Again we shall prove the contrapositive)

Assume $W(t_o, t_1)$ is singular for all $t_1 > t_o$. This implies there exists a vector $c(t_1)$ such that

$$(2\text{-}8) \quad c(t_1)\left[\int_{t_o}^{t_1} \Phi(t_o, t)\, H(t)\, H^T(t)\, \Phi^T(t_o, t)\, dt\right] c^T(t_1) \equiv 0$$

for any $t_1 \geq t_o$. From continuity of the integrand,

$$c(t_1)\, \Phi(t_o, t) H(t) H^T(t) \Phi^T(t_o, t) c^T(t_1) = 0 \qquad \text{for } t_o \leq t \leq t_1 .$$

Letting 0 denote a zero vector, it follows that

$$0 \equiv c(t_1)\, \Phi(t_o, t)\, H(t) \equiv c(t_1)\, \Phi^{-1}(t, t_o)\, H(t),$$

thus b, defined by $b(t) \equiv c(t_1)\, \Phi^{-1}(t, t_o)$ is an admissible vector in the sense that $b \cdot h = 0$ for all columns h of H, showing that b lies in the subspace spanned by the rows of B.

Define a scaler valued function

$$\Psi(t, x) \equiv c(t_1)\, \Phi^{-1}(t, t_o)\, x .$$

We will show that Ψ is an integral of the pfaffian equation associated with b, i.e., the equation

$$c(t_1)\, \Phi^{-1}(t, t_o) dx - (\, c(t_1)\, \Phi^{-1}(t, t_o) A(t) x)\, dt = 0.$$

Indeed $\Psi_x(t, x) = c(t_1)\, \Phi^{-1}(t, t_o)$, while $\Psi_t(t, x) = c(t_1)\, \dot{\Phi}^{-1}(t,t_o)x = - c(t_1)\, \Phi^{-1}(t, t_o)\, A(t)\, x$ which is as required.

Since the condition; rank $[G, AG, \ldots, A^{n-1}G \;\; = n]$, for the system (1.2) can be deduced from the more general criteria that $W(t_o, t_1)$ have an inverse for some $t_1 > t_o$, the verification that our extended criterion of complete controllability is equivalent with the existing criteria for linear systems, is completed.

Geometric Interpretation of Definition (2.2)

By associating a pfaffian system of the form (2-3) with the system (2-1), it is conspicuous that the stress is taken away from the functional form of the elements of the matrix H, and placed only on what the range of H(t, x), considered as an operator on E^r, is. This obviously should be the case when controls are required to be only finite valued and measurable.

In their paper [7], Markus and Lee consider a system of the form $\dot{x} = F(x, u)$, $F \in C^1$ in $E^n \times \Omega$, where Ω, a compact set contained in E^n with 0 in its interior, is the range set of the control. Assuming $F(0, 0) = 0$ and letting $A = F_x(0, 0)$, $H = F_u(0, 0)$, it is shown that if the linear system $\dot{x} = Ax + Hu$ is completely controllable, then the set of points from which the origin can be reached in finite time, by trajectories of $\dot{x} = F(x, u)$, is an open connected set containing the origin. Kalman [9] pointed out that a similar result can be obtained for a system of the form $\dot{x} = F(t, x, u)$ by assuming the linear approximation is completely oontrollable in terms of the criterion given in theorem 1.1.

We next proceed with an analysis, similar to that used in the papers mentioned above, to examine local controllability about a given trajectory of the system (2-1). Let $x(t_o) = 0$ be initial data for this system, v an arbitrary control (finite valued and measurable) and φ^v the corresponding solution. Let $u(t; \xi_1, \ldots, \xi_n) = u(t, \xi)$ be a family of controls such that $u(t, o) = v(t)$, u_ξ exists, and denote $x(t; \xi)$ as the response to $u(t; \xi)$. Then $x(t; \xi)$ satisfies

$$x(t; \xi) \equiv \int_{t_o}^{t} \left[g(\tau, x(\tau; \xi)) + H(\tau, x(\tau; \xi))\, u(\tau; \xi) \right] d\tau$$

$$x_\xi(t; 0) \equiv \int_{t_o}^{t} \Big[g_x(\tau, \varphi^v(\tau))\, x_\xi(\tau; 0) + H_x(\tau, \varphi^v(\tau)) v(\tau)$$

$$x_\xi(\tau; 0) + H(\tau, \varphi^v(\tau))\, u_\xi(\tau, 0) \Big] d\tau$$

where $H_x v$ is an n x n matrix with i jth element being $\sum_{\nu=1}^{r} H^{i\nu}_{x_j} v^\nu$.

For each $t \geq t_o$, we view $x(t; \xi)$ as a mapping $\xi \longrightarrow x$ with $0 \longrightarrow \varphi^v(t)$. Letting Z(t) denote the Jacobian matrix $x_\xi(t; 0)$, if it can be shown that for some $\bar{t}$, $Z(\bar{t})$ is non-singular, it follows that the attainable set of time $\bar{t}$, contains a neighborhood of the point $\varphi^v(\bar{t})$.

Let $\Phi(t, t_o)$ be a fundamental solution matrix of the system

$$\dot{x}(t) = \left[g_x(t, \varphi^v(t)) + H_x(t, \varphi^v(t))v(t) \right] x(t).$$

Then

$$Z(t) \equiv \int_{t_o}^{t} \Phi(t,\tau)\ H(\tau, \varphi^v(\tau))\ u_{\xi}(\tau; 0)\ d\tau\ .$$

From corollary 1.1 and theorem 1.1, a necessary and sufficient condition that there exist an $n \times r$ matrix u_{ξ} (t; 0) such that $Z(t_1)$ is non-singular for some $t_1 > t_o$, is that the linear system

$$(2\text{-}9) \qquad \dot{y}(t) = \left[g_x(t, \varphi^v(t)) + H_x(t, \varphi^v(t))v(t) \right] y(t) + H(t, \varphi^v(t))u(t)$$

be completely controllable. In terms of the pfaffian approach, let B(t, x) satisfy (2-2) while b(t, x) is again an arbitrary linear combination of rows of B. Then there exists an nxr matrix u_{ξ} (t, 0) such that $Z(t_1)$ is non-singular for some $t_1 > t_o$ iff the pfaffian system

$$B(t, \varphi^v(t))dx - B(t, \varphi^v(t)) \left[g_x(t, \varphi^v(t)) + H_x(t, \varphi^v(t))v(t) \right] x\,dt = 0$$

is non-integrable. From definition 2-1, this means

$$(2\text{-}10) \qquad b(t, \varphi^v(t))dx - b(t, \varphi^v(t)) \left[g_x(t, \varphi^v(t)) + H_x(t, \varphi^v(t))v(x) \right] x\,dt = 0$$

is not an exact differential, for arbitrary b.

It is interesting, at this point, to see the implications of the assumption that (2-10) *is* an exact differential. This implies and is implied by the condition

$$(2\text{-}11) \qquad \frac{d}{dt}\, b(t, \varphi^{v}(t)) \equiv -b(t, \varphi^{v}(t)) \left[g_{x}(t, \varphi^{v}(t)) + H_{x}(t, \varphi^{v}(t)) v(t) \right] ,$$

which can be recognized as the so-called adjoint system of the maximum principle approach to the time optimal problem for system (2-1). It should be noted that if $p(t) \equiv b(t, \varphi^{v}(t))$ satisfies (2-11), then it is an adjoint vector which is orthogonal to all of the columns of H. Since the maximum principle, for control components bounded by one in absolute value, implies: choose $u^{j}(t) \equiv + \operatorname{sgn} \sum_{i=1}^{n} p^{i}(t)\, H^{ij}(t, \varphi^{v}(t))$; in this case it yields no information since $b(t, x)\, H(t, x) \equiv 0$.

I shall designate such a problem as one which admits a *totally singular* arc, i.e., where the maximum principle yields no information in the time optimal problem, for any components of the optimal control. The problem would be singular, but not totally singular, if p is orthogonal to some, but not all columns of H.

I return again to the assumption that (2-10) is an exact differential. Since $b(t, x)\, H(t, x) \equiv 0$, for any vector $v(t)$, $0 \equiv \frac{\partial}{\partial x} \left[b(t,x) H(t,x) v(t)) \right]$ or $v(t)\, H^{T}(t, x)\, b_{x}(t, x) \equiv -b(t)\, H_{x}(t, x)\, v(t)$. Evaluating this identity at the point $(t, \varphi^{v}(t))$, substituting into (2-11) and expanding the left side yields

$$(2\text{-}12) \quad b_{t}(t, \varphi^{v}(t)) + b(t, \varphi^{v}(t))\, g_{x}(t, \varphi^{v}(t)) + g(t, \varphi^{v}(t))\, b_{x}^{T}(t, \varphi^{v}(t)) \equiv v(t)\, H^{T}(t, \varphi^{v}(t)) \left[b_{x}(t, \varphi^{v}(t)) - b_{x}^{T}(t, \varphi^{v}(t)) \right] .$$

The identity (2-12) is a necessary and sufficient condition that (2-10) be an exact differential.

Lemma 2.1 If the system (2-1) is not completely controllable at t_o, i.e., the pfaffian system (2-3) is integrable, then the matrix Z(t) is singular for all $t \geq t_o$, and for all reference trajectories φ^v.

Proof: If the pfaffian system (2-3) is integrable, then for some b, (2-4) is an exact differential. This implies and is implied by the conditions

$$b_t(t, x) \equiv -b(t, x)\, g_x(t, x) - g(t, x)\, b_x^T(t, x)$$

$$b_x(t, x) - b_x^T(t, x) \equiv 0.$$

Evaluating these two identities at $(t, \varphi^v(t))$ for an arbitrary control v, shows that (2-12) is satisfied, hence Z(t) is singular for all $t \geq t_o$. ▮

It should be stressed at this point, that it has not been shown that if for some control v, the matrix Z(t) is singular for all $t \geq t_o$, then sufficiently small n-nbds. of a point $\varphi^v(t)$ contain points not attainable in time t, from x_o at time t_o.

Another conjecture which one might be tempted to make is that if the pfaffian system (2-3) is not integrable, then (2-1) contains no totally singular arcs. This is not true, as the following example from [3] shows.

Example: $\dot{x}_1 = x_1^2 - x_1^2 x_2 u \qquad x_1(0) = 1$

$\dot{x}_2 = -x_2 + u \qquad x_2(0) = 0$.

For the time optimal problem of reaching the point (2, 0), it is shown in [3] that $u \equiv 0$ is the optimal (singular) control, if the restriction

$|u(t)| \leq 1$ is imposed, and it easily follows that this is also the optimal control in the class of finite valued measurable controls.

For this problem, one can use for the matrix B, the single vector $(1, x_1^2 x_2)$. The associated pfaffian equation is

$$dx_1 + x_1^2 x_2 \, dx_2 + x_1^2(x_2^2 - 1)dt = 0.$$

Let $x = (x_1, x_2)$,

$$a(x) = (1, x_1^2 x_2, x_1^2(x_2^2 - 1)).$$

Then $(\text{curl } a(x)). \, a(x) = 2 \, x_2 \, x_1^2 \not\equiv 0 \Longrightarrow$ the pfaffian is <u>not</u> <u>integrable</u>.

The optimal path from the point (1, 0) to $(\alpha, 0)$, $\alpha > 1$, is obtained with control $u \equiv 0$ and is

$$\varphi^0(t) \equiv \begin{cases} \dfrac{1}{1-t} \\ 0 \end{cases} .$$

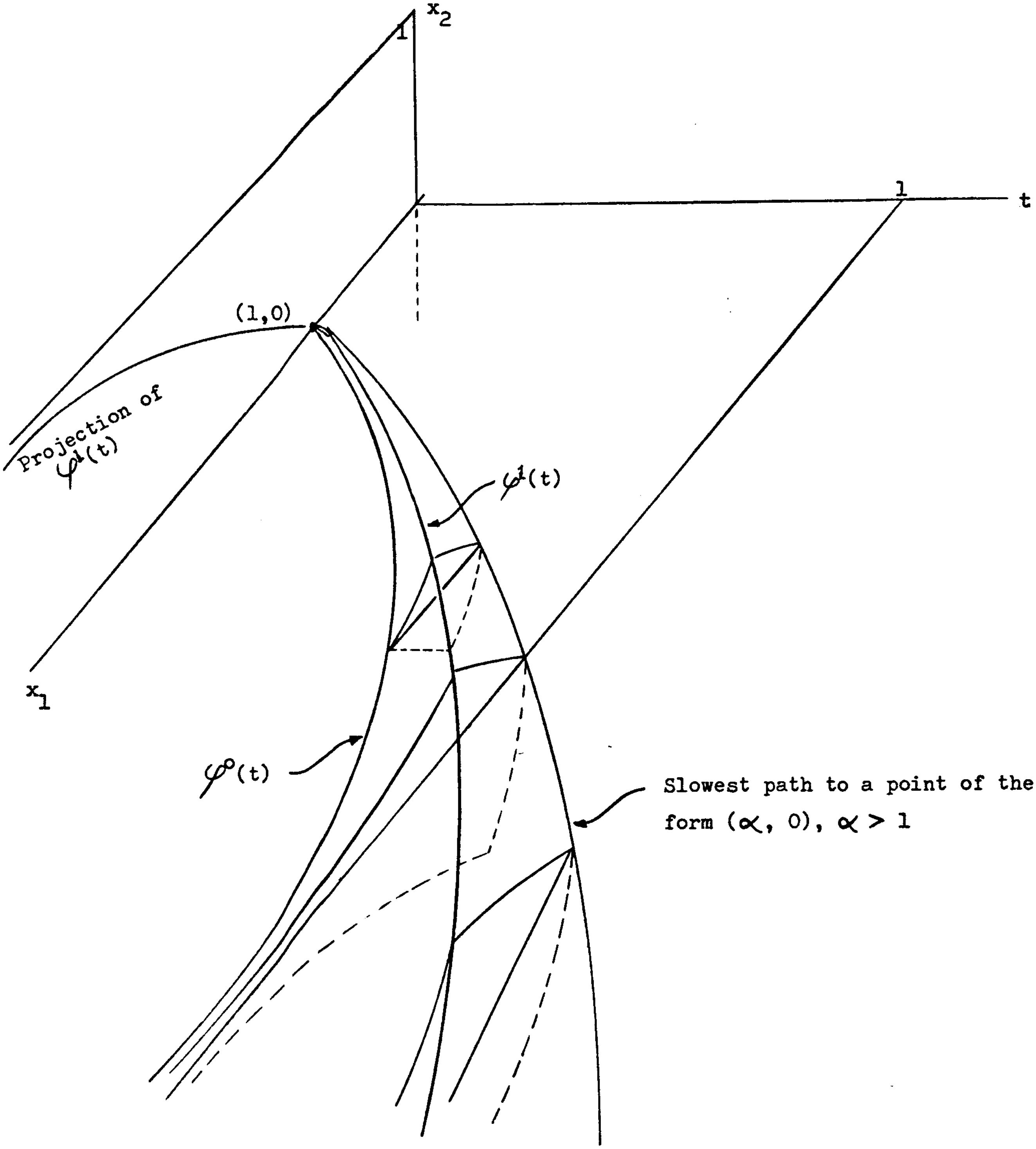
Thus $b(t, \varphi^0(t)) \equiv (1, 0)$. We next compute

$$b(t, \varphi^0(t)) \, . \, dx - b(t, \varphi^0(t)) \left[g_x(t, \varphi^0(t)) + H_x(t, \varphi^0(t)) \, . \, 0 \right] x \, d \, t$$

$$\equiv dx_1 + 0 \, dx_2 - \frac{2 \, x_1}{1 - t} \, dt \, .$$

Let $\bar{a}(x, t) \equiv (1, 0, \frac{-2x_1}{(1-t)})$. Then $(\text{curl } \bar{a}). \, \bar{a} \equiv 0$ which implies the pfaffian $dx_1 + 0 \, dx_2 - \frac{2x_1}{1-t} dt = 0$ is integrable, and the problem admits a totally singular arc. Pictured below is the reachable set, from (1, 0), with the control constraint $|u(t)| \leq 1$. Changing this constraint to

$|u(t)| \leq M$ does not essentially change the figure and in particular, does not change the arc $\varphi^{o}(t)$, which is such that all neighborhoods of a point $\varphi^{o}(t)$ contain points not attainable from (1, 0); even when the control class is chosen to be the class of finite valued, measurable functions.


x_2
1
1
t
(1,0)
Projection of $\varphi^{1}(t)$
x_1
$\varphi^{1}(t)$
$\varphi^{o}(t)$
Slowest path to a point of the form $(\alpha, 0)$, $\alpha > 1$

REFERENCES

1. Caratheodory, C., Untersuchungen urber die Grundlagen der Thermodynamik; Mathematische Ann. (1909), pp. 355-386.

2. Chow, W. L., Uber Systeme von Linear Particllen Differentialgleichungen erster Ordnung; Math. Ann. (1940), pp. 98-105.

3. Hermes, H., and Haynes, G., On the Nonlinear Control Problem with Control Appearing Linearly; SIAM Journal on Control, Ser. A, Vol 1, No 2, (1963), pp. 85-107.

4. Kalman, R. E., Contributions to the Theory of Optimal Control, Boletin De La Sociedad Mathematica Mexicana, 1960, pp. 102-119.

5. Kalman, R. E., Ho, Y. C., and Narendra, K. S., Controllability of Linear Dynamical Systems, Contributions to Diff. Eqs., Vol.I, No.2

6. LaSalle, J. P., The Time Optimal Control Problem; Contributions to the Theory of Nonlinear Oscillations, Vol 5.

7. Lee, E. B., and Markus, L., Optimal Control for Nonlinear Processes; Arch. for Rat. Mech. and Anal., Vol 8, No 1. 4.7 (1961) pp 36-58.

8. Nirenberg, L., Functional Analysis, New York University Lecture Notes, 1960, 1961.

PERTURBATION SOLUTIONS FOR LOW THRUST ROCKET TRAJECTORIES

D. P. JOHNSON
L. W. STUMPF

10 January 1964

AERONUTRONIC DIVISION
FORD MOTOR COMPANY

CONTRACT NO. NAS 8-5248

ABSTRACT

20965

A

Perturbation solutions of the equations of motion are presented which define the motion of a vehicle subjected to a low, constant thrust acceleration. A complete second-order solution is given for the case in which the thrust vector makes an arbitrary but constant angle with the radius vector. Application of the theory to transfers between circular orbits is discussed.

Author

PERTURBATION SOLUTIONS FOR LOW THRUST ROCKET TRAJECTORIES

INTRCDUCTION

Under Contract NAS 8-5248, Aeronutronic has been investigating the motion of a vehicle subjected to a low, constant thrust acceleration. The intent of tl.e investigation has been to improve the analytical representation of low thrust trajectories through perturbation solutions to the system equations of motion. Of primary interest is the application of the perturbation solutions to orbit transfer problems. By making use of these solutions, certain optimization problems of interest may be treated within the realm of simple optimization theory, and improved numerical computation techniques can be developed. This paper summarizes the perturbation theory, which includes a complete second-order theory for the motion where the thrust vector is maintained at a constant angle with respect to the radius vector. In subsequent sections we will introduce the system equations, present a first-order solution for tangential thrusting in order to illustrate the basic methodology, and then proceed to develop the general second-order theory. We will then conclude by discussing the application of the theory to orbit transfer problems using an energy/momentum approach.

EQUATIONS OF MOTION

Consider the problem of finding perturbation solutions of the differential equations of motion of a rocket moving under low thrust. The equations of motion are

$$(1) \qquad \frac{d^2\rho}{d\tau^2} - \frac{\nu^2}{\rho} + \frac{1}{\rho^2} = \alpha \cos \psi$$

$$(2) \qquad \frac{1}{\rho} \frac{d}{d\tau} (\rho\nu) = \alpha \sin \psi$$

The notation is that of E. Levin (Ref. 2). We define $\nu = \rho \frac{d\theta}{d\tau}$, α = (thrust acceleration ÷ initial gravitational acceleration), where $0 < \alpha < \frac{1}{8}$, ψ is the angle from the radius vector counterclockwise to the thrust vector while ρ, θ, τ are dimensionless position and time variables.. (See Figure 1)

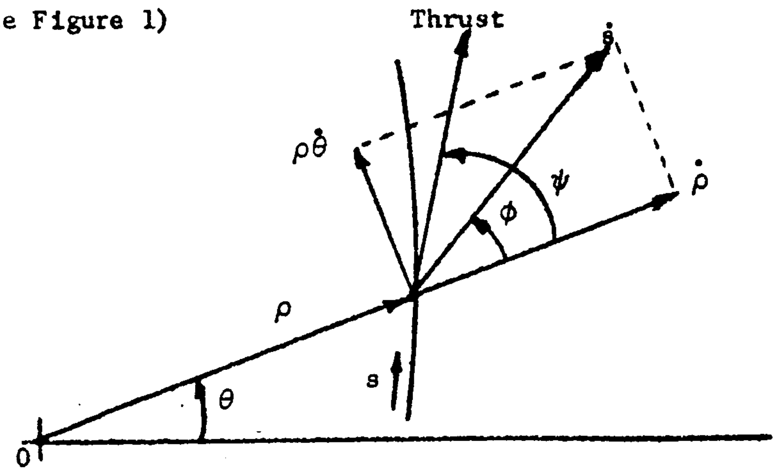


FIGURE 1

The constants of integration are determined according to the initial conditions, $\rho = 1, \dot{\rho} = 0, \tau = 0, \dot{\theta} = 1, \theta = 0$, $s = 0$, where s is a dimensionless arc length. The notation $\dot{\rho}$ and $\dot{\theta}$ denotes the differentiation of ρ and θ with respect to τ.

TANGENTIAL ACCELERATION

Consider the case of tangential acceleration: $\tan \psi = \frac{\nu}{\dot{\rho}}$, Then the equations of motion, (1) and (2), become

$$(3) \quad \frac{d^2\rho}{d\tau^2} - \frac{\nu^2}{\rho} + \frac{1}{\rho^2} = \frac{\alpha\dot{\rho}}{\sqrt{\dot{\rho}^2 + \nu^2}}$$

$$(4) \quad \frac{1}{\rho} \frac{d}{d\tau}(\rho\nu) = \frac{\alpha\nu}{\sqrt{\dot{\rho}^2 + \nu^2}}$$

With $(ds)^2 = (d\rho)^2 + \rho^2 (d\theta)^2$, the dimensionless arc length, and $\nu = \rho \frac{d\theta}{d\tau}$, we obtain

$$(5) \quad (\frac{ds}{d\tau})^2 = (\frac{d\rho}{d\tau})^2 + \nu^2 = V^2 .$$

We note from Levin (Ref. 2) that

$$(6) \quad \dot{E} = \alpha\dot{\rho}\cos\psi + \alpha\nu \sin\psi$$

where $\dot{E} = \frac{dE}{d\tau}$ the rate of change of the instantaneous energy E. For the tangential case, equation (6) becomes

$$(7) \quad \dot{E} = \alpha\dot{\rho}(\frac{\dot{\rho}}{V}) + \alpha\nu(\frac{\nu}{V}) = \alpha V$$

Now, changing to the independent variable, s, we have

$$V \frac{dE}{ds} = \alpha V$$

(8) $E = \alpha s + E_o$

Expressing equation (8) in terms of the speed V for an initially circular orbit, we arrive at the first integral obtained by Benney (Ref. 1)

(9) $$V^2 = \left(\frac{ds}{d\tau}\right)^2 = \frac{2}{\rho} + 2\alpha s - 1$$

Thus equation (4) can be written as

(10) $$\frac{d(\rho\nu)}{\rho\nu} = \frac{\alpha\, d\tau}{\sqrt{\dot{\rho}^2 + \nu^2}} = \frac{\alpha\, d\tau}{\frac{ds}{d\tau}} = \frac{\alpha ds}{\left(\frac{ds}{d\tau}\right)^2} = \frac{\alpha\, ds}{\frac{2}{\rho} + 2\alpha s - 1}$$

Integrating equation (10) we obtain

(11) $$h = \rho\nu = e^{\left\{\alpha\int \frac{ds}{\frac{2}{\rho} + 2\alpha s - 1}\right\}},$$

where $h = \rho\nu$ is the angular momentum.

With the approximation $e^x \approx 1 + x + \frac{x^2}{2}$, for small x, equation (11) reduces to

(12) $$h = \rho\nu = 1 + \alpha g(s) + \frac{\alpha^2}{2} g^2(s) ,$$

where $g(s) = \int \frac{ds}{\frac{2}{\rho} + 2\alpha s - 1}$

Similarly equation (3) can be written as

(13) $$\frac{d^2\rho}{d\tau^2} - \frac{\nu^2}{\rho} + \frac{1}{\rho^2} = \alpha \frac{d\rho}{ds}.$$

We next show that equations (3) and (4) can be brought into the form

$$(14)\qquad \frac{d^2u}{d\theta^2} + u = e^{\left\{-2\,\alpha\int \frac{ds}{2u+2\alpha\, s-1}\right\}}$$

$$(15)\qquad \frac{d\tau}{d\theta} = \frac{1}{u^2}\, e^{\left\{-\alpha\int \frac{ds}{2u+2\alpha\, s-1}\right\}}$$

where $u = \frac{1}{\rho}$. Setting $\rho = \frac{1}{u}$ in equation (11) we obtain

$$(16)\qquad \rho\nu = \rho^2\,\dot{\theta} = \rho^2\,\frac{d\,\theta}{d\,\tau} = \frac{\dot{\theta}}{u^2} = e^{\alpha g(s)} = \frac{1}{u^2}\,\frac{d\,\theta}{d\,\tau} = h$$

which is equation (15). Setting $\rho = \frac{1}{u}$ in equation (3) we get,

$$\frac{d\rho}{d\tau} = -\frac{1}{u^2}\,\frac{du}{d\theta}\,\frac{d\theta}{d\tau} = -h\,\frac{du}{d\theta}\,,\quad \frac{d^2\rho}{d\tau^2} = -h^2u^2\,\frac{d^2u}{d\theta^2} - h\,u^2\,\frac{dh}{d\theta}\,\frac{du}{d\theta}$$

Substituting these derivatives into equation (13) we obtain

$$(17)\qquad \frac{d^2u}{d\theta^2} + u = \frac{1}{h^2} - \frac{1}{h}\left[\frac{dh}{d\theta}\,\frac{du}{d\theta} + \frac{\alpha}{hu^2}\,\frac{d\rho}{ds}\right].$$

For the tangential case, we will show that the quantity inside the bracket is zero.

From equations (15), (16), and (9), it follows that

$$\frac{dh}{d\tau} = \frac{\alpha\, h}{\dot{s}^2}\,\frac{ds}{d\tau} = \frac{\alpha\, h}{\dot{s}} = \frac{dh}{d\theta}\,\dot{\theta} = h\,u^2\,\frac{dh}{d\theta} \qquad \text{or} \qquad \frac{dh}{d\theta} = \frac{\alpha}{u^2\dot{s}}$$

Also $\frac{d\rho}{ds} = -\frac{1}{u^2}\,\frac{du}{d\theta}\,\frac{d\theta}{ds}$

From equation (4) $\frac{d\theta}{ds} = \frac{u^2}{\alpha}\,\frac{dh}{d\tau} = \frac{h\,u^2}{\dot{s}}$

so that $\frac{d\dot{\rho}}{ds} = -\frac{h}{s}\frac{du}{d\theta}$.

Thus the quantity inside the bracket in equation (17) is zero, thereby substantiating equation (14).

We regard the right side of equation (14) as a function of θ so that we have a nonhomogeneous linear differential equation with constant coefficients. The complete solution is obtainable by variation of parameters in the form

$$(18) \quad u = A\cos\theta + B\sin\theta - \frac{i}{2}\, e^{i\theta}\int e^{-i\theta}\, {}^{-2\alpha\, g(s)}d\theta + \frac{i}{2}\, e^{-i\theta}\int e^{i\theta}\, {}^{-2\alpha\, g(s)}d\theta$$

where $i = \sqrt{-1}$, A and B are arbitrary constants and $g(s) = \int\frac{ds}{2u+2\alpha s-1}$

Integrating equation (18) by parts, we have

$$(19) \quad u = A\cos\theta + B\sin\theta \dot{+} e^{-2\alpha g(s)} + 0\,(\alpha^2)$$

as a first order solution. To obtain a more explicit form for the particular integral, we differentiate

$$(20) \quad u = e^{-2\alpha g(s)} = e^{-2\alpha\int\frac{ds}{2u+2\alpha s-1}}$$

with respect to s to obtain

$$(21) \quad \frac{du}{ds} = \frac{-2\alpha}{2u+2\alpha s-1}\, e^{-2\alpha\int\frac{ds}{2u+2\alpha s-1}} = \frac{-2\alpha u}{2u+2\alpha s-1}$$

which can be arranged as

$$(22) \quad u\ ds + s\ du = \frac{1}{2\alpha}\ du - \frac{1}{\alpha}\ u\ du = d(us)$$

which has the solution $us = \frac{u}{2\alpha} - \frac{u^2}{2\alpha} + c_1$, or since $u = 1$ when $s = 0$,

$$(23) \quad u = 1-2\alpha s.$$

Using equation (23) in equation (15) and knowing that

$$d\theta = \sqrt{u^2 - \frac{1}{u^2}\left(\frac{du}{ds}\right)^2}\ ds \qquad \text{we find that}$$

$$\theta = s + 0(\alpha) .$$

Now, evaluating the constants of integration in equation (19) according to the initial conditions, we get $A = 0$ and $B = 2\alpha$, so that equation (19) becomes

$$(24) \quad u = 1 - 2\alpha(s - \sin s) + 0(\alpha^2)$$

or since $\rho = \frac{1}{u}$ we have an expression which is in agreement with Benney's result.

$$(25) \quad \rho = 1 + 2\alpha(s - \sin s) + 0(\alpha^2)$$

We now write equation (15), neglecting terms of order α^2,

$$(26) \quad d\tau = [1 - 2\alpha(s - \sin s)]^{-2}(1 - \alpha s)\, d\theta + 0(\alpha^2)$$

$$= [1 + \alpha(3s - 4 \sin s)]\, d\theta$$

Using $d\theta = \sqrt{u^2 - \frac{1}{u^2}\left(\frac{du}{ds}\right)^2}\ ds$, we get

$$(27) \quad d\theta = [1 - 2\alpha(s - \sin s)]\ ds$$

We can integrate equation (26) to get

$$(28) \quad \tau - C = \int [1 + \alpha(3s - 4 \sin s)]\,[1 - 2\alpha(s - \sin s)]\ ds$$

$$= s + \alpha\left(\frac{s^2}{2} + 2 \cos s\right)$$

$s = 0$ when $\tau = 0$ so that $C = -2\alpha$ and equation (28) can be written as

$$(29) \quad \tau = s + \alpha\left(\frac{s^2}{2} - 4 \sin^2 \frac{s}{2}\right) .$$

Equations (25) and (29) constitute a complete first order solution of equations (14) and (15) in the case of tangential acceleration. Analogously, a second order solution can be derived.

GENERAL CASE

Let us now consider the more general case of thrusting with a constant orientation angle ψ. The equations of motion are now of the form of equations (1) and (2). Writing equation (6) in a different form, we have

$$(30) \quad \overset{\circ}{E} = V \frac{dE}{ds} = \alpha \cos(\phi - \psi) \sqrt{\dot{\rho}^2 + \nu^2}$$

or

$$(31) \quad E = \alpha \int \cos(\phi - \psi)\, ds + C = \frac{1}{2} V^2 - \frac{1}{\rho} = \frac{1}{2} \dot{s}^2 - \frac{1}{\rho} .$$

We choose $C = -\frac{1}{2}$ to satisfy initial conditions, $\rho = 1$, $s = 0$, $\tau = 0$, and then equation (31) becomes the first integral

$$(32) \quad V^2 = s^2 = \frac{2}{\rho} + 2\alpha f(s) - 1$$

where $f(s) = \int \cos(\phi - \psi)\, ds = s \sin\psi + 0(\alpha)$

In the same way that we obtained equation (11) we now obtain

$$(33) \quad h = \rho\nu = e^{\left\{\alpha \int \frac{\sin \psi}{\nu V} ds\right\}}$$

In place of equation (17) we will have

$$(34)\quad \frac{d^2u}{d\theta^2} + u = \frac{1}{h^2} - \frac{1}{h}\left[\frac{dh}{d\theta}\,\frac{du}{d\theta} + \frac{\alpha\cos\psi}{h\,u^2}\right] .$$

Since $h = \frac{1}{u^2}\,\frac{d\theta}{d\tau}$ and $\frac{dh}{d\theta} = \frac{\alpha\sin\psi}{uV}\sqrt{(\frac{d\rho}{d\theta})^2 + \rho^2}$, we can write equation (34) as

$$(35)\quad \frac{d^2u}{d\theta^2} + u = \frac{1}{h^2} - \frac{\alpha}{hV}\left[\frac{\sin\psi}{u}\cdot\frac{du}{d\theta}\sqrt{(\frac{d\rho}{d\theta})^2 + \rho^2} + \cos\ \frac{ds}{d\theta}\right] .$$

Hence we can write equations (35) and (33) as

$$(36)\quad \frac{d^2u}{d\theta^2} + u = \exp\left\{-2\alpha\int\frac{\sin\psi\,ds}{\nu V}\right\} - \frac{\alpha}{V}\left[\frac{1}{u}\ \sin\psi\sqrt{(\frac{d\rho}{d\theta})^2 + \rho^2}\ \frac{du}{d\theta}\right.$$

$$\left. + \cos\psi\ \frac{ds}{d\theta}\right]\quad \exp\left\{-\alpha\int\frac{\sin\psi\,ds}{\nu V}\right\} ,$$

$$(37)\quad \frac{d\tau}{d\theta} = \frac{1}{u^2}\quad \exp\left\{-\alpha\int\frac{\sin\psi\,ds}{\nu V}\right\} ,\ \text{where } e^x \equiv \exp\{x\}$$

The complete solution of equation (36) is obtainable by variation of parameters, analogously to equation (18), as

$$(38)\quad u = A\cos\theta + B\sin\theta - \frac{i}{2}\,e^{i\theta}\int e^{-i\theta}\,F_1(\theta)\,d\theta + \frac{i}{2}\,e^{-i\theta}\int e^{i\theta}\,F_1(\theta)d\theta$$

where

$$F_1(\theta) = \exp\left\{-2\int\frac{\alpha\sin\psi}{\nu V}\ \frac{ds}{d\theta}\ d\theta\right\} - \frac{\alpha}{V}\left[\frac{1}{u}\ \sin\psi\sqrt{(\frac{d\rho}{d\theta})^2 + \rho^2}\ \frac{du}{d\theta}\right.$$

$$\left. + \cos\psi\,\frac{ds}{d\theta}\right]\qquad \exp\left\{-\int\frac{\alpha\sin\psi}{\nu V}\ \frac{ds}{d\theta}\ d\theta\right\}$$

Using the approximation $e^x \approx 1 + x + \frac{x^2}{2}$ for small x, we obtain a first order expression for equation (38).

$$(39) \quad u = A \cos \theta + B \sin \theta - \frac{i}{2} e^{i\theta} \int e^{-i\theta} F_2(\theta) d\theta + \frac{i}{2} e^{-i\theta} \int e^{i\theta} F_2(\theta) d\theta$$

where $F_2(\theta) = 1 - 2\alpha \int \frac{\sin \psi}{\nu V} \frac{ds}{d\theta} \cdot d\theta - \frac{\alpha}{V} \cos \psi \frac{ds}{d\theta} + 0(\alpha^2)$.

Upon integrating equation (39) and evaluating the constants of integration, A and B, according to the initial conditions $\theta = 0$, $u = 1$, $\frac{du}{d\theta} = 0$, and letting $\psi = \psi_o$, a constant, we obtain

$$(40) \quad u = 1 - \alpha [2 \sin \psi_o (\theta - \sin \theta) + \cos \psi_o (1 - \cos \theta)] + 0(\alpha^2).$$

$$(41) \quad \rho = 1 + \alpha [2 \sin \psi_o (\theta - \sin \theta) + \cos \psi_o (1 - \cos \theta)] + 0(\alpha^2).$$

Using equation (37) it follows that

$$(42) \quad \tau = \theta + \alpha [\sin \psi_o (\frac{3}{2} \theta^2 + 4 \cos \theta - 4) + 2 \cos \psi_o (\theta - \sin \theta)] + 0(\alpha^2).$$

Since $(ds)^2 = (d\rho)^2 + \rho^2 (d\theta)^2$, we can write $ds = \rho d\theta + 0(\alpha^2)$ so that by equation (41)

$$(43) \quad s = \theta + \alpha [\sin \psi_o (\theta^2 + 2 \cos \theta - 2) + \cos \psi_o (\theta - \sin \theta)] + 0(\alpha^2).$$

In general, the basic solution to equations (1) and (2) can be written in any of the two explicit forms.

$$(44) \quad \rho = \rho(\theta), \quad \tau = \tau(\theta)$$

$$(45) \quad \rho = \rho(\tau), \quad \theta = \theta(\tau)$$

We have already shown that equations (41) and (42) take the form of (44). From equations (41) and (42) we obtain the form (45), namely,

$$(46) \quad \rho = 1 + \alpha [2 \sin \psi_o (\tau - \sin \tau) + \cos \psi_o (1 - \cos \tau)] + 0(\alpha^2)$$

(47) $\theta = \tau - \alpha\left[\sin\psi_o\left(\frac{3}{2}\tau^2 + 4\cos\tau - 4\right) + 2\cos\psi_o(\tau - \sin\tau)\right] + 0(\alpha^2).$

Other relations involving s are

(48) $\theta = s - \alpha\left[\sin\psi_o(s^2 + 2\cos s - 2) + \cos\psi_o(s - \sin s)\right] + 0(\alpha^2),$

(49) $\rho = 1 + \alpha\left[2\sin\psi_o(s - \sin s) + \cos\psi_o(1 - \cos s)\right] + 0(\alpha^2),$

(50) $\tau = s + \alpha\left[\sin\psi_o\left(\frac{1}{2}s^2 + 2\cos s - 2\right) + \cos\psi_o(s - \sin s)\right] + 0(\alpha^2).$

We will now develop second-order expressions in the two explicit forms as noted by (44) and (45). From our earlier first-order results we obtain the following expressions:

$$u = 1 - \alpha[2\sin\psi_o(\theta - \sin\theta) + \cos\psi_o(1 - \cos\theta)] + 0(\alpha^2)$$

$$\frac{du}{d\theta} = \alpha[-2\sin\psi_o(1 - \cos\theta) - \cos\psi_o\sin\theta] + 0(\alpha^2)$$

$$\rho = 1 + \alpha[2\sin\psi_o(\theta - \sin\theta) + \cos\psi_o(1 - \cos\theta)] + 0(\alpha^2)$$

$$= \frac{ds}{d\theta} + 0(\alpha^2)$$

$$\frac{d\tau}{d\theta} = 1 + \alpha[\sin\psi_o(3\theta - 4\sin\theta) + 2\cos\psi_o(1 - \cos\theta)] + 0(\alpha^2)$$

$$\nu = 1 - \alpha[\sin\psi_o(\theta - 2\sin\theta) + \cos\psi_o(1 - \cos\theta)] + 0(\alpha^2)$$

$$= V + 0(\alpha^2).$$

Upon substituting these expressions into equation (38) and integrating and evaluating the constants of integration, A and B, according to the initial conditions we have

$$
\begin{aligned}
(51)\quad u = 1 + \alpha\,[\cos\psi_o\,(\cos\theta - 1) - 2\sin\psi_o\,(\theta - \sin\theta)\,] \\
+\,\alpha^2\,[\sin^2\psi_o\,(-\,2\theta^2 - 7\theta\sin\theta - 18\cos\theta + 18) \\
+\cos^2\psi_o\,(\theta\sin\theta + 2\cos\theta - 2) \\
+\sin\psi_o\cos\psi_o\,(-\,8\theta - \tfrac{11}{2}\,\theta\cos\theta + \tfrac{27}{2}\sin\theta)\,] + 0(\alpha^3)
\end{aligned}
$$

Since $(1 + \alpha x + \alpha^2 y)^{-1} = 1 - \alpha x + \alpha^2 (x^2 - y) + 0(\alpha^3)$ we can write equation (51) as

$$
\begin{aligned}
(52)\quad \rho = 1 + \alpha\,[\cos\psi_o\,(1 - \cos\theta) + 2\sin\psi_o\,(\theta - \sin\theta)\,] \\
+\,\alpha^2\,[\sin^2\psi_o\,(6\theta^2 + 4\sin^2\theta - \theta\sin\theta + 18\cos\theta - 18) \\
+\cos^2\psi_o\,(\cos^2\theta - \theta\sin\theta - 4\cos\theta + 3) \\
+\sin\psi_o\cos\psi_o\,(12\theta + \tfrac{3}{2}\,\theta\cos\theta - \tfrac{19}{2}\sin\theta + 4\sin\theta\cos\theta\,] + 0(\alpha^3).
\end{aligned}
$$

Upon substituting equation (47) into equation (52) we obtain ρ in terms of τ, namely,

$$
\begin{aligned}
(53)\quad \rho = 1 + \alpha[\cos\psi_o(1 - \cos\tau) + 2\sin\psi_o\,(\tau - \sin\tau)\,] \\
+\,\alpha^2\,[\sin^2\psi_o\,(3\tau^2 + 3\tau^2\cos\tau + 5\cos^2\tau + 6\cos\tau + 2\tau\sin\tau - 11) \\
+\sin\psi_o\cos\psi_o\,(8\tau + \tfrac{11}{2}\,\tau\cos\tau - \tfrac{3}{2}\,\tau^2\sin\tau - \tfrac{19}{2}\sin\tau \\
-\,4\sin\tau\cos\tau) + \sin^2\tau - 3\tau\sin\tau - 4\cos\tau + 4\,] + 0(\alpha^3).
\end{aligned}
$$

We obtain τ in terms of θ from equation (37)

(54) $$\tau = \theta + \alpha[2 \cos \psi_o (\theta - \sin \theta) + \sin \psi_o (\frac{3}{2} \theta^2 + 4 \cos \theta - 4)]$$
$$+ \alpha^2 [\sin^2 \psi_o (\frac{7}{3} \theta^3 + 24 \sin \theta + 6 \theta \cos \theta - 6 \sin \theta \cos \theta - 24 \theta)$$
$$+ \cos^2 \psi_o (2 \theta \cos \theta - 12 \sin \theta + \frac{3}{2} \sin \theta \cos \theta + \frac{17}{2} \theta)$$
$$+ \sin \psi_o \cos \psi_o (\frac{23}{2} \theta^2 + 37 \cos \theta + \theta \sin \theta - 6 \cos^2 \theta - 31)] + 0(\alpha^3).$$

From equation (54) and using equation (47) we obtain θ in terms of τ

(55) $$\theta = \tau + \alpha [\sin \psi_o (- \frac{3}{2} \tau^2 - 4 \cos \tau + 4) - 2 \cos \psi_o (\tau - \sin \tau)]$$
$$+ \alpha^2 [\sin^2 \psi_o (\tau^3 - 8 \sin \tau + 6 \tau \cos \tau - 6 \tau^2 \sin \tau - 10 \sin \tau \cos \tau + 12 \tau)$$
$$+ \cos^2 \psi_o (8 \sin \tau - 6 \tau \cos \tau + \frac{5}{2} \sin \tau \cos \tau - \frac{9}{2} \tau)$$
$$+ \sin \psi_o \cos \psi_o (- \frac{5}{2} \tau^2 - 21 \cos \tau - 15 \tau \sin \tau - 10 \cos^2 \tau - 3 \tau^2 \cos \tau + 31)]$$
$$+ 0(\alpha^3)$$

We now have enough information available to obtain energy and momentum expressions. Using equations (31), (32), and (33) we have

(57) $$E = - \frac{1}{2} + \alpha \tau \sin \psi_o$$
$$+ \alpha^2 [\sin^2 \psi_o (- \frac{1}{2} \tau^2 - \cos \tau + 1) + \sin \psi_o \cos \psi_o (\tau - \sin \tau$$
$$+ 8 \sin \tau \cos \tau) - \cos \tau + 1] + 0(\alpha^3)$$

(58) $$h = 1 + \alpha \tau \sin \psi_o + \alpha^2 [\sin^2 \psi_o (\tau^2 + 2 \cos \tau - 2) + \sin \psi_o \cos \psi_o (\tau - \sin \tau)]$$
$$+ 0(\alpha^3)$$

The previously developed equations constitute second-order solutions to equations (1) and (2) for $0 < \alpha < \frac{1}{8}$, ψ_o a constant, and θ, τ, and s not too large.

ORBIT TRANSFER PROBLEMS

The above theory may be applied almost directly to the problem of transfer from one circular orbit to another. For transfers involving an initial thrusting phase on departure from a circular orbit followed by a coasting phase and a subsequent thrusting phase to establish the final circular orbit, it is only necessary to choose the thrusting angles and thrust durations so that the energy and momentum values at the beginning and end of the coasting phase are identical. Figure 2 illustrates a sample transfer.

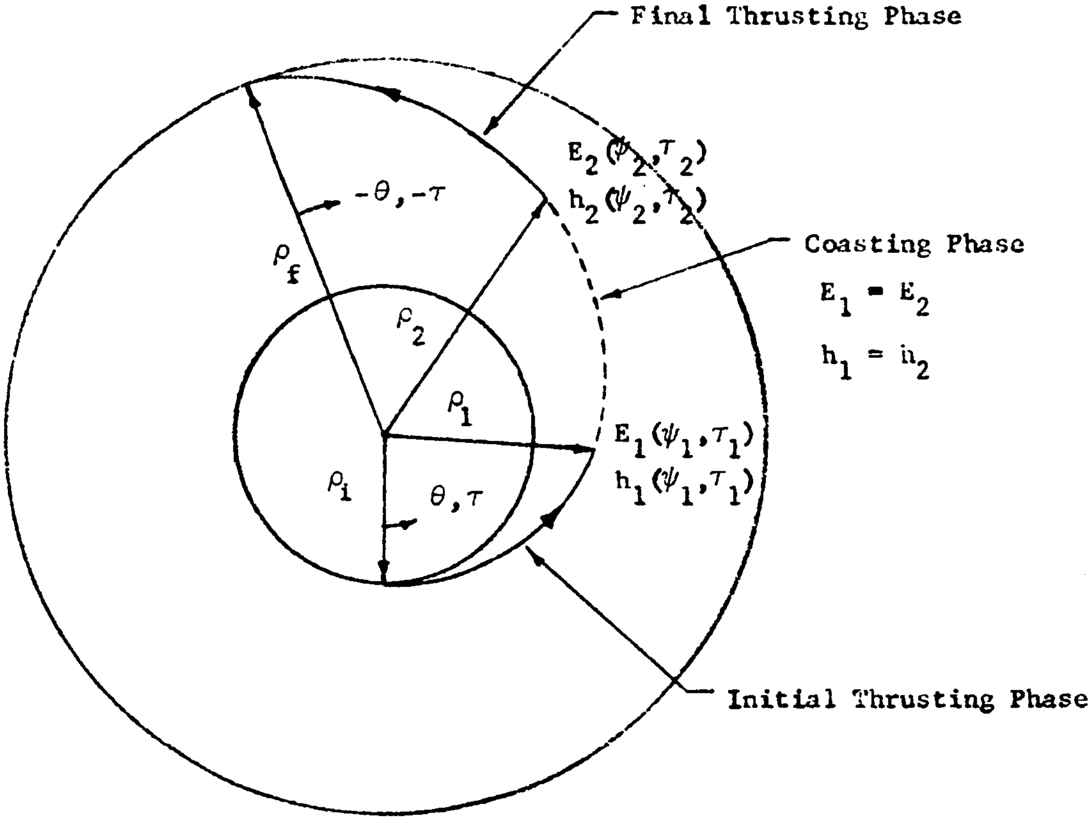


FIGURE 2

The final boundary condition of the second thrusting phase may be satisfied by interpreting it as an initial condition and then considering the motion in negative time. Because the various equations were non-dimensionalized with regard to the initial orbit, it is necessary to examine the conversion factors between quantities measured in the $\rho_i = 1$ system and the $\rho_f = 1$ system. Denoting the ρ_f quantities by primes, and defining $K = \frac{\rho_f}{\rho_i}$, the conversion factors are:

($\rho_i = 1$ quantity)	x	conv. factor	=	($\rho_f = 1$ quantity)
E		K		E'
h		$K^{-\frac{1}{2}}$		h'
α		K^2		α'
τ		$K^{-\frac{3}{2}}$		τ'
length		K^{-1}		length'

When these conversion factors are utilized, the non-dimensional equations are sufficient to define the transfer maneuver.

REFERENCES

1. Benney, D. J., "Escape From a Circular Orbit Using Tangential Thrust," Jet Propulsion, March 1958, pp. 167-169.

2. Levin, E., "Lunar and Interplanetary Trajectories," Handbook of Astronautical Engineering, ed. H. H. Koelle, Mc Graw-Hill Book Company Inc., New York, 1961, chap. 9-1.

REPUBLIC AVIATION CORPORATION

CORRECTIONS TO PROGRESS REPORT NO. 4
"APPLICATION OF THE TWO FIXED CENTER PROBLEM TO LUNAR TRAJECTORIES"

By

Mary Payne

Farmingdale, L. I., N. Y.

CORRECTIONS TO PROGRESS REPORT NO. 4 "APPLICATION OF THE TWO FIXED CENTER PROBLEM TO LUNAR TRAJECTORIES"

By

Mary Payne

The conclusions reached from the old data were slightly different from those based on the newer data. The corrected conclusions are to be found on the first page of these corrections. This page corrects the conclusions found on pages 236 and 237 in Progress Report No. 4.

The data presented on pages 248 and 249, Tables II and III, in Progress Report No. 4 are to be replaced by the tables presented here.

The conclusions on the relative merits of the various methods are to be revised as follows:

A and C are best for long range on the return leg

B and C have a slight superiority for midcourse

D is best in moon reference, on the first leg and for short range on the return leg.

E and F are inferior almost everywhere

It may be noted that the deviations in moon reference are approximately 100 times as large as for corresponding deviations in earth reference. This is perhaps to be expected since the ratio of earth to moon mass is 80, and hence the terms neglected in moon reference should be approximately 80 times as large as those neglected in earth reference.

Table II. Moon Reference From Earth to Moon
α = .987883194 Corresponds to Center of Rotation at the Moon

		A	B	C	D	E	F	α
59-60	Δx	-4.8	4.2	-3.5	.38	-1.1	.77	.74542049
	Δy	-17	13	-13	.20	-1.2	.69	
	Δz	-5.6	4.3	-4.2	.02	-1.3	.08	
59-66	Δx	-179	297	-160	67	-219	485	.95560202
	Δy	-849	842	-811	16	-705	648	
	Δz	-278	267	-268	.45	-119	63	
59-71	Δx	-213	838	-211	447		13728	.98774102
	Δy	-2018	1982	-2018	-129		-3695	
	Δz	-715	625	-715	-118		-6390	
66-71	Δx	-13	58	-12	24	-2390	6815	.98774026
	Δy	-409	368	-409	-24	-2795	-8514	
	Δz	-151	130	-151	-12	-395	-5567	

Table III. Moon Reference Moon Towards Earth
$\alpha = .987883194$ Corresponds to Center of Rotation at the Moon

		A	B	C	D	E	F	α
72-73	Δx	-2.6	3.8	-2.5	.66	-3930	-3930	.98746055
	Δy	-18	18	-18	.02	119	119	
	Δz	-5.9	5.7	-5.9	-.10	227	227	
72-75	Δx	-21	46	-18	14	-14509	-14509	.98750712
	Δy	-160	151	-158	-3.6	-439	-439	
	Δz	-53	46	-53	-3.3	1463	1463	
72-84	Δx	-12	828	48	424	-53692	-53677	.98751836
	Δy	-2607	2015	-2579	-295	1277	1277	
	Δz	-883	581	-877	-151	17063	17050	
75-80	Δx	-26	73	-7.0	33	-19552	-19552	.97489278
	Δy	-457	184	-454	-135	497	496	
	Δz	-153	138	-158	-10	1146	1146	
75-84	Δx	-13	234	62	147	-36241	-36237	.97552707
	Δy	-1500	1371	-1497	-62	822	822	
	Δz	-523	435	-529	-46	2605	2605	
80-84	Δx	-3.8	30	9.3	20	-15278	-15277	.84840941
	Δy	-296	446	-231	108	-169	-170	
	Δz	-100	76	-91	-7.7	487	487	

NASA TM X-53024 APPROVAL PAGE March 17, 1964

PROGRESS REPORT NO. 5
on Studies in the Fields of
SPACE FLIGHT AND GUIDANCE THEORY

Sponsored by Aero-Astrodynamics Laboratory of
Marshall Space Flight Center

The information in this report has been reviewed for security classification. Review of any information concerning Department of Defense or Atomic Energy Commission programs has been made by the MSFC Security Classification Officer. This report, in its entirety, has been determined to be unclassified.

E. D. GEISSLER, Director
Aero-Astrodynamics Laboratory

DISTRIBUTION

DIR, Mr. F. Williams

R-FP, Dr. Ruppe

I-I/IB-DIR, Col. James

R-P&VE, Mr. Swanson
Mr. Moore
Mr. Richard
Mr. Gassaway
Mr. Taylor
Mr. Brooks
Mr. Hosenthien
Mr. Scofield
Mr. Woods
Mr. Digesu
Mr. R. Hill
Dr. R. Decher
Mrs. Neighbors

R-COMP, Dr. Arenstorf
Mr. Davidson
Mr. Harton
Mr. Schollard
Mr. Seely
Mr. Calhoun

R-AERO, Dr. Geissler
Dr. Hoelker
Dr. Speer
Mr. Miner (80)
Mr. Braud
Mr. Schmieder
Mr. Dearman
Mr. Schwaniger
Mr. Ingram
Mr. G. Herring
Mr. Powers
Mr. Callaway
Mr. Thomae
Mr. Baker

R-AERO, Mr. Hart
Mr. Lovingood
Mr. Winch
Mr. Tucker
Mr. Hooper
Mrs. Chandler
Mr. Lisle
Mr. Kurtz
Mr. Fine
Mr. Stone
Mr. deFries
Mr. Teague
Dr. Sperling
Dr. Heybey
Mr. Cummings
Mr. Jean
Mr. Felker

R-RP-U, Mr. Bland

AST-S, Dr. Lange

R-DIR, Mr. Weidner
Dr. McCall

MS-IP, Mr. Ziak

MS-IPL (8)

MS-H

CC-P

Mr. Hans K. Hinz (4)
Research Department
Grumman Aircraft Engineering Corporation
Bethpage, Long Island, New York

Mr. J. S. Farrior
Lockheed
P. O. Box 1103
West Station
Huntsville, Alabama

Dr. George Nomicos (5)
Applied Mathematics Section
Applied Research and Development
Republic Aviation Corporation
Farmingdale, Long Island, New York

Mr. Theodore N. Edelbaum
Senior Research Engineer
Research Laboratories
United Aircraft Corporation
East Hartford, Connecticut

Mr. Arthur C. Gilbert, Sc. D.
Chief, Space Systems Requirements
Corporate Systems Center
Division United Aircraft Corporation
1690 New Britain Avenue
Farmington, Connecticut

Mr. Robert A. Lerman
Sr. Analytical Engineer
Technical Planning
Hamilton Standard Division
United Aircraft Corporation
Windsor Locks, Connecticut

Dr. M. G. Boyce (3)
Department of Mathematics
Vanderbilt University
Nashville, Tennessee

Dr. Daniel Dupree, Project Leader (15)
Department of Mathematics
Northeast Louisiana State College
Monroe, Louisiana

Dr. Steve Hu (10)
Northrop Corporation
Box 1484
Huntsville, Alabama

Dr. William Nesline, Jr. (3)
Analytical Research Department
Missile and Space Division
Raytheon Company
Bedford, Massachusetts

Dr. Robert W. Hunt
Department of Mathematics
Southern Illinois University
Carbondale, Illinois

Mr. Robert Silber
Department of Mathematics
Southern Illinois University
Carbondale, Illinois

Dr. H. Hermes
Mail #A127
Martin Company
P. O. Box 179
Denver 1, Colorado

Mr. R. W. Reck (5)
Martin Company
3313 S. Memorial Parkway
Huntsville, Alabama

Dr. M. L. Anthony
Space Flight Technology
Mail No. A-153
The Martin Company
Denver 1, Colorado

Mr. Samuel Pines (10)
Analytical Mechanics Associates, Inc.
941 Front Street
Uniondale, New York

Dr. W. A. Shaw (10)
Mechanical Engineering Department
Auburn University
Auburn, Alabama

Mr. J. W. Hanson (20)
Computation Center
University of North Carolina
Chapel Hill, North Carolina

Mr. Oliver C. Collins
Advanced Research
Flight Technology Department
Aero-Space Division
Organization 2-5762
Mail Stop 15-12
Boeing Company
P. O. Box 3707
Seattle, Washington

Mr. Carl B. Cox
Chief, Flight Technology Dept.
Advanced Research
Aero-Space Division
Organization 2-5762
Mail Stop 15-03
Boeing Company
P. O. Box 3707
Seattle 24, Washington

Mr. Richard Hardy
Box AB-38
Boeing Company
P. O. Box 1680
Huntsville, Alabama

Mr. Wes Morgan
Box AB-49
Boeing Company
P. O. Box 1680
Huntsville, Alabama

Mr. Robert Glasson (2)
Bendix Systems Division
Bendix Corporation
3322 Memorial Parkway South
Huntsville, Alabama

Chrysler Corporation Missile Division
Sixteen Mile Road and Van Dyke
P. O. Box 2628
Detroit 31, Michigan
ATTN: Mr. T. L. Campbell (2)
Dept. 7162
Applied Mathematics

Mr. Harry Passmore (3)
Hayes International Corporation
P. O. Box 2287
Birmingham, Alabama

Astrodynamics Operation
Space Sciences Laboratory
Missile and Space Vehicle Department
General Electric Company
Valley Forge Space Technology Center
P. O. Box 8555
Philadelphia 1, Pennsylvania
ATTN: Mr. J. P. deVries
Dr. V. Szebehely
Mr. C. Cavoti

Dr. I. E. Perlin
Rich Computer Center
Georgia Institute of Technology
Atlanta, Georgia

Dr. Charles C. Conley
Department of Mathematics
University of Wisconsin
Madison, Wisconsin

Dr. O. R. Ainsworth
Department of Mathematics
University of Alabama
Huntsville, Alabama

Mr. D. Johnson
Astrodynamics Section
Astrosciences Department
Aeronutronic Div. of Ford Motor Co.
Ford Road
Newport Beach, California

Dr. N. N. Puri
Electric Engineering Department
Drexel Institute of Technology
Philadelphia, Pennsylvania

University of Kentucky
College of Arts and Sciences
Department of Mathematics and Astronomy
Lexington, Kentucky
ATTN: Dr. J. C. Eaves (4)

Space Sciences Laboratory
Space and Information Systems Division
North American Aviation, Inc.
Downey, California
ATTN: Dr. D. F. Bender

Mr. H. A. McCarty
Space and Information Systems Division
North American Aviation, Inc.
12214 Lakewood Blvd.
Downey, California

Mr. J. W. Scheuch
P. O. Box 557
North American Aviation, Inc.
Huntsville, Alabama

Mr. Joe Mason
Space and Information Systems Division
Department 41-595-720
North American Aviation, Inc.
12214 Lakewood Blvd.
Downey, California

Dr. D. M. Schrello
Director, Flight Sciences
Space and Information Systems Division
North American Aviation, Inc.
12214 Lakewood Blvd.
Downey, California

Mr. Myron Schall
Supervisor, Powered Trajectory
Space and Information Division
North American Aviation, Inc.
12214 Lakewood Blvd.
Downey, California

Mr. S. E. Cooper
Space and Information Division
Dept. 41-697-610
North American Aviation, Inc.
12214 Lakewood Blvd.
Downey, California

Space Sciences Laboratory
Space and Information Systems Division
North American Aviation, Inc.
12214 Lakewood Blvd.
Downey, California
ATTN: Mr. Paul DesJardins
Mr. Harold Bell

Dr. E. R. Rang
Military Products Group
Aeronautical Division
Minneapolis-Honeywell Regulator Co.
Mail Stop #340
Minneapolis, Minnesota

Mr. Dahlard Lukes
M.P.G. - Aeronautical Division
Minneapolis-Honeywell Regulator Co.
2600 Ridgway Road
Minneapolis, Minnesota

Dr. Robert M. L. Baker, Jr.
Astrodynamics Research Center
Lockheed-California Company
Burbank, California

Mr. Ralph Q. Haumacher
A2-863: Space/Guidance & Control
Douglas Aircraft Corporation
3000 Ocean Park Blvd.
Santa Monica, California

Mr. F. A. Hewlett
Manager of Documentation
Federal Systems Division
IBM
6702 Gulf Freeway
Houston 17, Texas

Mr. George Cherry
Massachusetts Institute of Technology
Cambridge, Massachusetts

Mr. William C. Marshall
Reseach Engineer
M. P. G. Research Dept. - Sta #340
Minneapolis-Honeywell Regulator Co.
2600 Ridgway Road
Minneapolis 3, Minnesota

Mr. W. G. Melbourne
Jet Propulsion Laboratory
4800 Oak Grove Drive
Pasadena 3, California

Dr. Byron D. Tapley
Department of Aerospace Engineering
University of Texas
Austin, Texas

Dr. Hans G. Baussus
Aerospace Group
General Precision
McBride Avenue
Little Falls, New Jersey

Dr. William A. Mersman
Chief,
Electronic Machine Computing Branch
Ames Research Center
Moffett Field, California

Siegfried J. Gerathewohl
Chief,
Biotechnology Division
Ames Research Center
Moffett Field, California

Dr. Herman M. Dusek
A. C. Research and Development
A. C. Spark Plug
The Electronics Division
of General Motors
950 N. Sepulveda Blvd.
El Segundo, California

Mr. Howard S. London
Rm. 910-E
Bellcomm, Inc.
1100 17th Street, N. W.
Washington, D. C.

Mr. Howard Haglund
Jet Propulsion Laboratory
4800 Oak Grove Drive
Pasadena 3, California

Mr. Hewitt Phillips
Langley Research Center
Hampton, Virginia

Mr. Frank J. Carroll
Systems Requirement Department
Equipment Division
Raytheon Company
40 Second Avenue
Waltham, Massachusetts

Mr. T. Perkins
Chrysler Corporation
HIC Building
Huntsville, Alabama

Mr. Robert Allen
Manager, Huntsville Sales Office
A. C. Spark Plug
The Electronics Division
of General Motors
Holiday Office Center
Huntsville, Alabama

Mr. Dale B. Ruhmel
Staff Associate
Aeronautical Research Associates of Princeton, Inc.
50 Washington Road
Princeton, New Jersey

Dr. Raymond Rishel
Mathematical Analysis Staff
Organization 2-5330
Mail Stop 89-75
Boeing Company
P. O. Box 3707
Seattle, Washington

Dr. Rudolf Hermann
Director
University of Alabama Research Institute
4701 University Avenue, N.W.
Huntsville, Alabama

Dr. S. H. Lehnigk
Physical Sciences Laboratory
Army Missile Command, Bldg. 5429
Redstone Arsenal, Alabama

Mr. R. J. Hayes
Code RE-TG
NASA Headquarters
Washington, D. C.

Mr. Harold Chestnut
1 River Road
Schenectady, New York

Space Sciences Laboratory
Space and Information Systems
North American Aviation, Inc.
Downey, California
ATTN: Mr. Dave Engles

Dr. Kirk Brouwer
Yale University Observatory
Box 2023, Yale Station
New Haven, Connecticut

Dr. Imre Izsak
Smithsonian Institution Astrophysical Observatory
60 Garden Street
Cambridge 38, Massachusetts

Dr. Peter Musen
Goddard Space Flight Center
NASA
Greenbelt, Maryland

Dr. Yoshihide Kozai
Smithsonian Institution Astrophysical Observatory
60 Garden Street
Cambridge 38, Massachusetts

Dr. Rudolph Kalman
Research Institute for Advanced Study
7212 Bellona Avenue
Baltimore 12, Maryland

Mr. Ken Kissel
Aeronautical Systems Division
Applied Mathematics Research Branch
Wright-Patterson Air Force Base
Dayton, Ohio

Mr. Jack Funk
Manned Spacecraft Center
Flight Dynamics Branch
NASA
Houston, Texas

Dr. J. B. Rosser
Department of Mathematics
Cornell University
Ithaca, New York

Douglas Aircraft Corporation
3000 Ocean Park Blvd.
Santa Monica, California
ATTN: R. E. Holmen A2-263
Guidance & Control Section

Dr. Joseph F. Shea
Office of Manned Space Flight
National Aeronautics and Space Administration
801 19th Street, N. W.
Washington 25, D. C.

Mr. Charles F. Pontious
Guidance & Navigation Program
Office of Advanced Research & Technology
Code: REG
National Aeronautics and Space Administration
Washington 25, D. C.

Mr. Jules Kanter
Guidance & Navigation Program
Office of Advanced Research & Technology
Code: REG
National Aeronautics and Space Administration
Washington 25, D. C.

Dr. Ray Wilson
OART-Code RRA
Washington 25, D. C.

Dr. Joseph W. Siry
Theory & Analysis Office (547)
Data Systems Division
Goddard Space Flight Center
Greenbelt, Maryland

Douglas Aircraft Corporation
3000 Ocean Park Blvd.
Santa Monica, California
ATTN: Mr. Joe Santa
A2-863

J. B. Cruz, Jr.
Research Associate Professor
Coordinated Science Laboratory
Urbana, Illinois

Mr. Alan L. Friedlander
Research Engineer
Guidance and Control Section
IIT Research Center
10 W. 35th Street
Chicago 16, Illinois

Mr. Joseph V. Natrella
Manned Space Flight - Code MCR
NASA Headquarters, FOB 10B
Washington, D. C.

Mr. Donald Jezewski
Guidance Analysis Branch
Spacecraft Technology Division
Manned Spacecraft Center
Houston, Texas

Mr. Robert M. Williams
Chief, Guidance Analysis
General Dynamics/Astronautcs
Mail Zone 513-0
P. O. Box 166
San Diego 12, California

Mr. M. D. Anderson, Jr.
General Dynamics Corporation
Suite 42 Holiday Office Center
South Memorial Parkway
Huntsville, Alabama

Mr. Y. L. Luke
Mathematics & Physics Division
Midwest Research Institute
425 Volker Boulevard
Kansas City 10, Missouri

Mr. Ted Guinn
Advanced Space Technology
Engineering Research
Douglas Aircraft Corporation
Santa Monica, California

Mr. Robert Chilton
NASA Manned Spacecraft Center - EG
P. O. Box 1537
Houston, Texas

Dr. S. E. Ross
Office of Manned Space Flight - NASA
Washington, D. C.

Stephen J. Kahne
Lt. USAF
Applied Mathematics Branch
Data Sciences Laboratory
Air Force Cambridge Research Laboratories
Office of Aerospace Research
Lawrence G. Hanscom Field
Bedford, Massachusetts

Mr. Daniel B. Killeen
Computer Laboratory
Norman Mayer Bldg.
Tulane University
New Orleans, Louisiana

Dr. Bernard Friedland
Staff Scientist - Control
General Precision, Inc.
Little Falls, New Jersey

Mr. Walter L. Portugal
Manager, Systems Sales
Aerospace Group
General Precision, Inc.
Little Falls, New Jersey

Mr. C. H. Tross
Manager, Aerospace Sciences
UNIVAC Division of Sperry Rand Corporation
P. O. Box 6248
San Diego 6, California

Mr. Ken Squires
Goddard Space Flight Center, Bldg. #1
National Aeronautics and Space Administration
Greenbelt, Maryland

Dr. Paul Degrarabedian
Astro Science Laboratory
Building G
Space Technology Laboratory, Inc.
One Space Park
Redondo Beach, California

Dr. George Leitmann
Associate Professor, Engineering Science
University of California
Berkeley, California

Dr. R. P. Agnew
Department of Mathematics
Cornell University
Ithaca, New York

Dr. Jurgen Moser
Professor of Mathematics
Graduate School of Arts and Science
New York University
New York, New York

Dr. Lu Ting
Department of Applied Mechanics
Polytechnic Institute of Brooklyn
333 Jay Street
Brooklyn 1, New York

Mr. Clint Pine
Department of Mathematics
Northwestern State College
Natchitoches, Louisiana

Dr. John Gates
Jet Propulsion Laboratory
4800 Oak Grove Drive
Pasadena 3, California

Auburn Research Foundation (2)
Auburn University
Auburn, Alabama

Grumman Library
Grumman Aircraft Engineering Corp.
Bethpage, L. I., New York

Jet Propulsion Laboratory
Library
4800 Oak Grove Drive
Pasadena 3, California

Scientific and Technical Information Facility (25)
ATTN: NASA Representative (A-AK/RKT)
P. O. Box 5700
Bethesda, Maryland

NASA Ames Research Center (2)
Mountain View, California
ATTN: Librarian

NASA Flight Research Center (2)
Edwards Air Force Base, California
ATTN: Librarian

NASA Goddard Space Flight Center (2)
Greenbelt, Maryland
ATTN: Librarian

NASA Langley Research Center (2)
Hampton, Virginia
ATTN: Librarian

NASA Launch Operations Directorate (2)
Cape Canaveral, Florida
ATTN: Librarian

NASA Lewis Research Center (2)
Cleveland, Ohio
ATTN: Librarian

NASA Manned Spacecraft Center (2)
Houston 1, Texas
ATTN: Librarian

NASA Wallops Space Flight Station (2)
Wallops Island, Virginia
ATTN: Librarian

Space Flight Library (4)
University of Kentucky
Lexington, Kentucky

University of Kentucky Library (10)
University of Kentucky
Lexington, Kentucky

Office of Manned Space Flight
NASA Headquarters
Federal Office Building #6
Washington 25, D. C.
ATTN: Mr. Eldon Hall
Dr. A. J. Kelley

Mr. Bryan F. Dorlin
Theoretical Guidance & Control Branch
NASA-Ames Research Center
Moffett Field, California

Lt. Col. R. A. Newman
Air Force Space Systems Div
M-SSVH
Bldg. 5250
Redstone Arsenal, Alabama